2013
2014

TFG

CARACTERIZACIÓN DEL COMPORTAMIENTO SUICIDA CON MAPAS AUTOORGANIZADOS

Autora: Marta Ramos Martín
Tutor: José Miguel Leiva Murillo

Ingeniería de Sistemas Audiovisuales
Universidad Carlos III de Madrid

ÍNDICE

"Lo único que puedo decirles a ustedes es que si me hacen una pregunta y no sé la respuesta, les diré directamente que no sé la respuesta, pero también buscaré la forma de encontrarla y, cuando la tenga, se la daré."

1. Introducción

El estudio de variables involucradas en el comportamiento suicida es importante desde un punto de vista médico, social y económico, si bien, la conducta suicida es resultado de una compleja interacción entre factores de vulnerabilidad y eventos medioambientales, haciendo más difícil la prevención o detección de intentos de suicidio. Numerosos factores de riesgo y protección han sido consistentemente identificados, pero los modelos de predicción para el comportamiento suicida son imprecisos. Comprendiendo mejor la jerarquía y organización de las variables relacionadas con la conducta suicida podrá mejorarse la detección de sujetos potenciales.

Ante la complejidad del problema, los métodos de estadística clásica no son capaces de trabajar con grandes muestras, elevados números de variables e interacciones fuertemente no lineales de los datos. Por otro lado, tanto el aprendizaje máquina moderno como la minería de datos proporcionan métodos que superan estas limitaciones y han sido aplicados con éxito en problemas de biología computacional como la clasificación de intentos de suicidio [Leiva-Murillo et al. (2012)].

Uno de los problemas estudiados con mayor intensidad en aprendizaje máquina es la selección de variables, que aplicado al objetivo del presente proyecto consiste en la identificación de los factores de riesgo más importantes en el comportamiento suicida. En aprendizaje supervisado[1], las variables de interés deben ser seleccionadas de acuerdo a su capacidad de predicción, lo que en términos de esta aplicación se refiere a la identificación de la presencia o no de conducta suicida. El ranking de variables es una simplificación del problema de selección de variables, y consiste en la disposición de las mismas en un orden de relevancia decreciente. Un ejemplo de la estadística clásica es el discriminante de Fisher [Fukunaga (1990)], método aplicado en este proyecto para la clasificación de variables.

Para esta aplicación, se propone el uso de Mapas Autoorganizados o SOM como método de identificación de factores relevantes debido a la relación no lineal existente entre las variables y el comportamiento suicida [Kohonen (2001)]. Existen varias razones por las que los SOM resultan útiles en la realización de esta tarea. En primer lugar, proporcionan un soporte visual que facilita el reconocimiento de la estructura de datos. Segundo, son capaces de trabajar con datos de elevadas dimensiones y patrones no lineales y, por último, aunque se pensase inicialmente en los SOM como método para aprendizaje no supervisado[2], su aplicación en problemas supervisados ha sido satisfactoria [Leiva-Murillo et al. (2012)].

[1] El aprendizaje supervisado es una técnica aplicada en aprendizaje automático y minería de datos basada en el aprendizaje guiado por una variable auxiliar.

[2] El aprendizaje no supervisado es un método de aprendizaje automático en el que un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori.

El presente estudio analiza una cohorte de 8.699 sujetos compilados por cinco grupos de investigación, procedentes de cuatro países diferentes y miembros del consorcio EURECA (*European Research Consortium for Suicide*): Montpellier, Geneva, Molise, Oviedo y Madrid. Cada sujeto ha sido caracterizado por 606 variables relacionadas con su entorno sociodemográfico, así como por sus respuestas a cuestionarios normalizados que determinan la hostilidad, impulsividad, alcoholismo, traumas de infancia, desesperación, etc.

Para la consecución del proyecto, se ha utilizado el paquete SOM Toolbox como herramienta de desarrollo, que incluye una serie de funciones para la implementación y visualización de Mapas Autoorganizados. La herramienta SOM Toolbox se ejecuta sobre la plataforma de simulación matemática Matlab, en su versión R2011a.

En los siguientes apartados se describirán aspectos relacionados con la solución propuesta a este problema. En el apartado 1.1. se indicará la importancia a nivel mundial de los intentos de suicidio, las estrategias puestas en práctica para su prevención y factores influyentes sobre la conducta suicida. En la sección 1.2. se explicará la estructura, uso y funcionamiento de los Mapas Autoorganizados. Por otro lado, en el apartado 1.3. se detallará el contenido y estructura del diccionario de variables del consorcio EURECA y en la sección 1.4. se describirá la herramienta de desarrollo del programa. El apartado 1.5. recoge los tres métodos discriminantes utilizados en el proyecto, mientras que en los apartados 1.6. y 1.7. se especificará, respectivamente, el marco legal y entorno socioeconómico relativos a los intentos de suicidio. Por último, en la sección 1.8. se desarrollarán los objetivos y principales motivaciones del proyecto.

1.1. Los Intentos de Suicidio

La Organización Mundial de la Salud (*World Health Organization*) estima que cada 3 segundos se produce un intento de suicidio. Cada año, se producen entre 10 y 20 millones de intentos de suicidio y, al menos, un millón de suicidios llegan a efectuarse [WHO (1999)]. Los datos son aún más alarmantes entre los grupos de jóvenes, ya que el suicidio supone un 6,3% de las causas de muerte en personas de entre 10 y 24 años [Leiva-Murillo et al. (2012)]. Estudios recientes determinan que los intentos de suicidio son de 10 a 40 veces más frecuentes que los suicidios consumados [López-Castromán et al. (2010)].

Las estrategias actuales para la prevención del suicidio están enfocadas principalmente tanto en la detección como en el tratamiento de trastornos mentales. Sin embargo, a pesar de los esfuerzos de prevención que incluyen mejoras en el tratamiento de la depresión, la prevalencia de los intentos de suicidio se ha mantenido sin cambios durante la última década. Esto sugiere que existe una necesidad de mejorar la comprensión de los factores de riesgo en los intentos de suicidio más allá de trastornos psiquiátricos [Ruiz et al. (2012)].

De acuerdo a la Estrategia Nacional para la Prevención del Suicidio (*National Strategy for Suicide Prevention*), un primer paso importante es identificar a aquellos sujetos con mayor riesgo de suicidio [Ruiz et al. (2012)]. Se ha demostrado que los intentos de suicidio son, de hecho, el mejor predictor para posteriores intentos de suicidio con un riesgo elevado de desenlace fatal [López-Castromán et al. (2010)].

Un análisis sistemático demuestra que las tasas de repeticiones de intentos de suicidio sin un final fatal son del 16% en el primer año, 23% del cuarto año en adelante y del 40% entre el tercer y octavo año de seguimiento [López-Castromán et al. (2010)].

Por otro lado, existen evidencias de que los genes juegan un papel importante en la predisposición a la conducta suicida. Estudios de adopción, gemelos o mellizos y familias demuestran que aproximadamente el 40% de variabilidad en el comportamiento suicida puede tener una base genética. Sin embargo, el efecto de los genes sobre comportamientos complejos como trastornos psiquiátricos o conductas suicidas es muy reducido [Baca-García et al. (2009)].

Hasta la fecha, no existen directrices universalmente aceptadas para tratar los intentos de suicidio de modo que puedan evitarse tras al alta médica, lo que puede deberse, en parte, al desconocimiento de la jerarquía o relación entre los factores que predicen la repetición de intentos de suicidio. En este grupo, se incluyen factores sociodemográficos como la edad, el desempleo, no estar casado o un bajo nivel de educación, así como factores clínicos tales como letalidad del primer intento de suicidio, salud física deficiente y enfermedades mentales [López-Castromán et al. (2010)].

Al mismo tiempo, la evaluación de factores de riesgo es imprecisa debido a que los modelos predictivos disponibles son también imprecisos, obstaculizando la tarea de identificación de repeticiones de intentos de suicidio [López Castromán et al. (2010)].

Los sujetos que cometen intentos de suicidio son extensamente reconocidos como un grupo de alto riesgo. La identificación precisa de sujetos con alto riesgo de conducta suicida es crucial para el desarrollo y la aplicación de intervenciones específicas de acuerdo con la posibilidad de reintento de suicidio y los recursos disponibles [López-Castromán et al. (2010)].

1.2. Mapas Autoorganizados

1.2.1. ¿Qué son los Mapas Autoorganizados?

Los Mapas Autoorganizados o SOM (*Self-Organizing Maps*) fueron descritos por primera vez en 1982 por el destacado académico e investigador finlandés Teuvo Kohonen como una red neuronal artificial, de ahí que el nombre que en ocasiones reciben sea redes o mapas de Kohonen [Kohonen (1982)].

Los SOM representan la proyección de los datos de estudio sobre una red bidimensional de celdas, que puede seguir un patrón cuadrangular o hexagonal, siendo este último el más común. Un modelo $\boldsymbol{m}_i$ es asociado a cada celda del mapa, tal y como se muestra en la Figura [1], y equivale a un vector de las mismas dimensiones que los datos. Cada elemento de los datos de estudio será mapeado en la celda cuyo modelo sea el más similar al de dicho elemento. [Kohonen and Honkela (2007)].

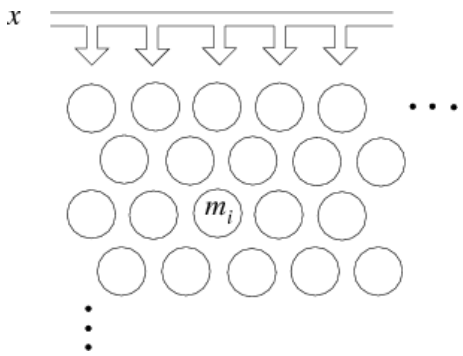


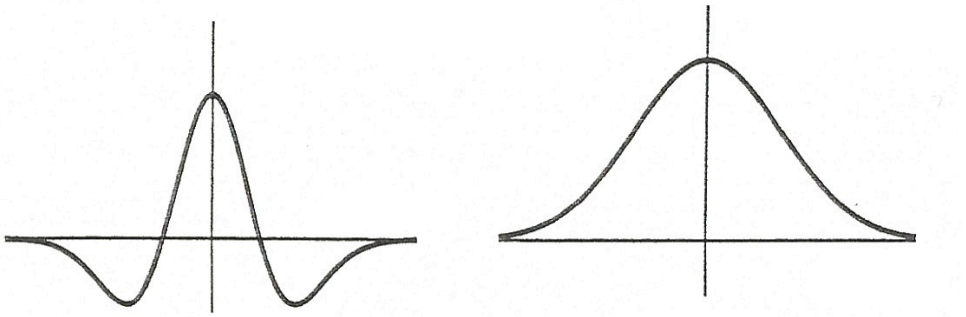
Figura [1] – Conjunto de nodos (neuronas) en un SOM bidimensional.

Cuando estos modelos se determinan aplicando el algoritmo del SOM, se observa que existe una mayor similitud entre celdas vecinas que entre celdas más alejadas. De esta manera, puede considerarse que el conjunto de modelos constituye una similitud gráfica y estructural de la distribución de los datos de estudio [Kohonen and Honkela (2007)].

El algoritmo SOM surgió de los modelos de redes neuronales tempranas, en particular, de los modelos de memoria asociativa y el aprendizaje adaptativo. Una nueva motivación surgió ante la necesidad de explicar la organización espacial de las funciones del cerebro, como se observa especialmente en la corteza cerebral. No obstante, el SOM no era el primer paso en esta dirección; ya existían los detectores de línea espacialmente ordenados de Malsburg (1973) y el modelo de campo neuronal de Amari (1980). Sin embargo, la capacidad de autoorganización de estos modelos era más bien débil [Kohonen and Honkela (2007)].

La intención de Kohonen era introducir un modelo de sistema compuesto por, al menos, dos subsistemas de diferente naturaleza. Uno de estos subsistemas se corresponde con una red neuronal competitiva que implementa la función WTA (*winner-takes-all*), un principio aplicado en modelos computacionales de redes neuronales por el que las neuronas de una misma capa compiten con otras neuronas para activarse. Sólo la neurona con el mayor grado de activación permanecerá activa mientras las demás se mantendrán inactivas [Kohonen and Honkela (2007)].

El segundo subsistema está controlado por la red neuronal y modifica la plasticidad sináptica local de las neuronas en el aprendizaje [Kohonen and Honkela (2007)]. Las sinapsis son las conexiones neuronales en el cerebro humano. Las conexiones sinápticas entre neuronas no son estáticas, sino que sufren modificaciones como consecuencia de una actividad o experiencia previa. Así, los estímulos del exterior pueden provocar que algunas sinapsis se potencien, mientras que otras se debiliten. Este proceso de plasticidad sináptica resulta esencial para el aprendizaje y la memoria [Investigación y Ciencia (2012)]. El subsistema de control de la plasticidad define cómo la actividad local determina el aprendizaje en la vecindad. Este subsistema es no negativo y puede adoptar una distribución Gaussiana. En la Figura [2] se representan los dos subsistemas de interacción: el subsistema de control de actividad, también conocido como *Mexican Hat Function* (a), y el subsistema de control de plasticidad (b) [Kohonen (2001)].

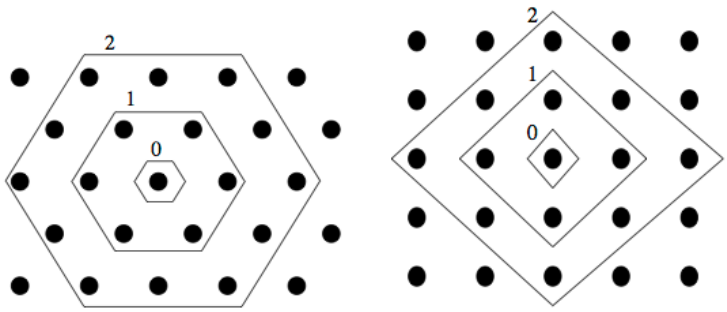


(a) Control de Actividad (b) Control de Plasticidad

Figura [2] – Subsistemas de interacción del modelo de Kohonen.

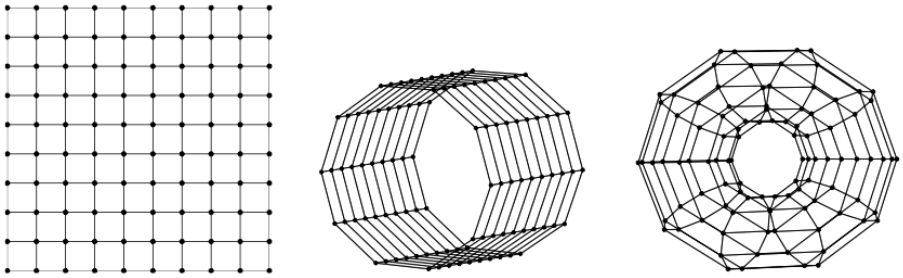
1.2.2. Algorítmica y Funcionamiento

Un SOM es un conjunto de neuronas organizadas en celdas regulares. Cada neurona está representada por un vector d-dimensional $\boldsymbol{m} = [m_1, \dots, m_d]$, donde d es igual a la dimensión de los vectores de entrada. Las neuronas se encuentran conectadas a neuronas adyacentes por una relación de vecindad, que determina la topología o la estructura del mapa . Los diferentes tipos de estructuras y formas se muestran en las Figuras [3] y [4] respectivamente.



(a) Celda Hexagonal (b) Celda Rectangular

Figura [3] – Estructuras de los vecindarios de dimensiones 0, 1 y 2.



(a) Plano (b) Cilindro (c) Toroide

Figura [4] – Diferentes formas de mapa.

El algoritmo de entrenamiento del SOM guarda cierta similitud con los algoritmos de cuantificación vectorial, como es el caso del algoritmo de las *k-medias*, cuyo objetivo es la partición de un conjunto de *n* observaciones en *k* grupos de modo que cada observación se asocie al grupo más cercano a la media. La principal diferencia es que, además del vector BMU (*Best-Matching Unit*), también se actualiza la topología del mapa: la región alrededor del vector BMU se estira hacia la actual muestra de entrenamiento, tal y como se observa en la Figura [5]. El resultado final es que las neuronas aparecen ordenadas, es decir, las neuronas más próximas entre sí estarán caracterizadas por vectores muy similares [Vesanto et al. (2000)].

El SOM es entrenado iterativamente. En cada paso de entrenamiento, se selecciona aleatoriamente una muestra del vector **x** de los datos de entrada y se calcula su distancia Euclídea con respecto al resto de vectores. La neurona cuyo vector sea el más cercano al vector de entrada **x** se denomina BMU (*Best-Matching Unit*) [Vesanto et al. (2000)]. Esta relación se expresa mediante la Ecuación [1] donde el BMU queda representado con el subíndice *c*.

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \qquad [1]$$

Una vez encontrado el BMU, los vectores del SOM son actualizados para que el BMU se sitúe próximo al vector de entrada **x**. Este procedimiento de adaptación, estira el BMU hacia el vector de muestra tal y como se representa en la Figura [5].

La regla de actualización del SOM para el vector asociado a la muestra *i* queda reflejada en la Ecuación [2].

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t)\left[x(t) - m_i(t)\right] \qquad [2]$$

donde t hace referencia al tiempo, el parámetro $\boldsymbol{x}(t)$ es uno de los vectores de entrada seleccionado aleatoriamente en el instante t, $h_{ci}(t)$ es el núcleo del vecindario alrededor del BMU denotado con el subíndice c, y $\alpha(t)$ es la tasa de aprendizaje en el instante t. En las Figuras [6] y [7] se representan gráficamente los parámetros $h_{ci}(t)$ y $\alpha(t)$.

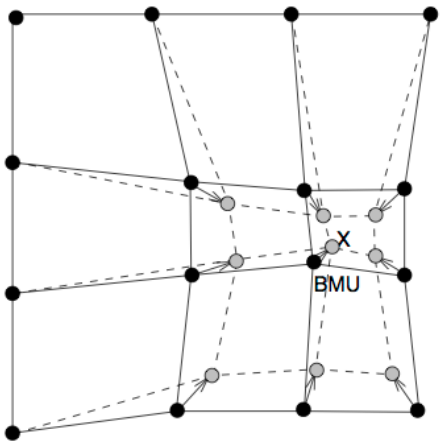


Figura [5] – Proceso de actualización del BMU (*Best-Matching Unit*).

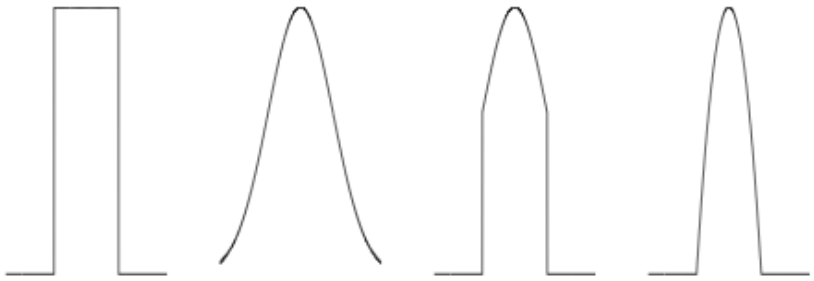


Figura [6] – Diferentes tipos de vecindario en función del valor de $h_{ci}(t)$.

(1) $h_{ci}(t) = \beta(\sigma_t - d_{ci})$ [3]

(2) $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2}$ [4]

(3) $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2}\beta(\sigma_t - d_{ci})$ [5]

(4) $h_{ci}(t) = \max\{0, 1-(\sigma_t - d_{ci})^2\}$ [6]

donde α_t es el radio del vecindario en el instante t, $d_{ci} = \|\boldsymbol{r}_c - \boldsymbol{r}_i\|$ es la distancia entre c y la celda i-ésima del mapa y $\boldsymbol{\beta}(x)$ es la función escalón, tal que:

$$\beta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad [7]$$

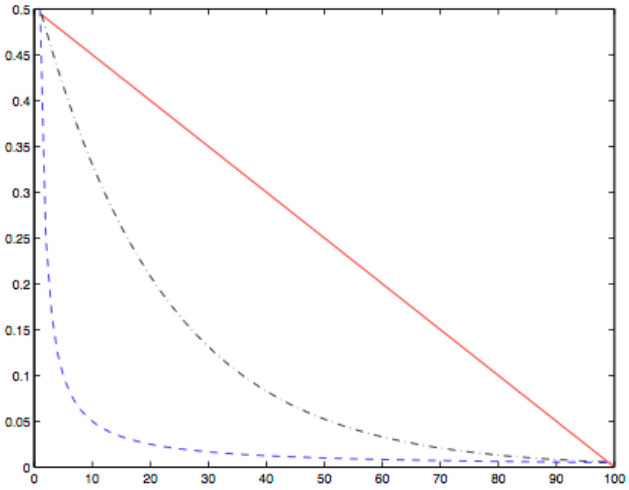


Figura [7] – Diferentes tasas de aprendizaje en función del valor de $\alpha(t)$.

$\alpha(t) = \alpha_0(1 - t/T)$ [8]

$\alpha(t) = \alpha_0(0.005/\alpha_0)^{t/T}$ [9]

$\alpha(t) = \alpha_0/(1 + 100\,t/T)$ [10]

donde T es la longitud de entrenamiento y α_0 es la tasa inicial de aprendizaje. Para este proyecto se ha aplicado un valor de núcleo de vecindario $h_{ci}(t)$ definido por la Ecuación [4] y una tasa de aprendizaje $\alpha(t)$ especificado en la Ecuación [9].

1.3. Diccionario EURECA

El consorcio EURECA ha reclutado 3.839 intentos de suicidio y suicidios consumados en los últimos años. Los datos clínicos y sociodemográficos de los sujetos de estudio han sido reunidos en una base de datos común junto a los resultados obtenidos mediante métodos de valoración desarrollados por equipos clínicos [Leiva-Murillo et al. (2012)] .

Todos los sujetos que habían manifestado un intento de suicidio fueron hospitalizados de acuerdo a lo que se define como: "un comportamiento potencialmente auto lesivo con un desenlace fatal para el que hay evidencias (tanto explícitas como implícitas) de que la persona está destinada con un nivel (distinto de cero) al suicidio". Esta definición ha sido adoptada por el NIMH (*National Institute of Mental Health*) y los principales grupos de investigación en la UE [Leiva-Murillo et al. (2012)].

Los estudios fueron aprobados por el Comité de Ética de la Investigación y realizados de acuerdo a los principios de la Declaración de Helsinki. Todos los participantes firmaron un formulario de consentimiento después de la explicación del objetivo y procedimientos del estudio.

Además de sujetos con historial suicida, la base de datos incluye pacientes psiquiátricos sin antecedentes en intentos de suicidio, donantes de sangre y pacientes ortopédicos sin antecedentes en trastornos mentales o comportamientos suicidas pero emparentados con individuos que han llevado a cabo un intento de suicidio.

Para todos los sujetos de estudio, se recopilaron también rasgos sociodemográficos. Los diagnósticos psiquiátricos fueron evaluados mediante las pruebas DIGS (*Diagnostic Interview for Genetics Studies*) y MINI (*Mini International Neuropsychiatric Interview*).

La conducta suicida ha sido evaluada haciendo uso de las escalas SIS (*Suicidal Intent Scale*) y RRRS (*Risk Rescue Rating Scale*). Los motivos por los que puede desarrollarse un comportamiento suicida fueron examinados con la escala SSI (*Scale for Suicidal Ideation*).

Las siguientes escalas validadas en diferentes idiomas fueron utilizadas para investigar las medidas intermedias en la conducta suicida: LHA (*Life History of Agression*), BDHI (*Buss-Durkee Hostility Inventory*), STAXI (*Spielberg State-Trait Anger Expression Inventory*), BIS10 (*Barratt Impulsivity Scale*), BDI (*Beck Depression Inventory*) y BHS (*Beck Hopelessness Scale*).

El cuestionario CTQ (*Childhoom Trauma Questionnaire*), una medida retrospectiva sobre el abuso de menores, y el cuestionario CAGE, relativo a los problemas con el alcohol, también están incluidos en la base de datos.

1.4. Entorno de Desarrollo SOM Toolbox

Como plataforma de desarrollo para el proyecto se ha recurrido a la herramienta SOM Toolbox para Matlab 5. Este paquete de funciones surgió como soporte para una óptima y sencilla implementación de Mapas Autoorganizados en Matlab con finalidades de investigación. En particular, los responsables del desarrollo de la SOM Toolbox trabajaron en el campo de la minería de datos, por lo que esta herramienta está orientada hacia esa dirección en forma de funciones de visualización de gran alcance [Vesanto et al. (2000)].

La plataforma SOM Toolbox puede utilizarse para el preprocesamiento de datos, inicialización y entrenamiento de Mapas Autoorganizados haciendo uso de un rango de diferentes topologías, visualizar SOM en distintos modos y analizar las propiedades de los SOM y los datos, como por ejemplo, la calidad del SOM, las celdas del mapa y las correlaciones entre variables [Vesanto et al. (2000)].

Existen dos tipos de funciones incluidas en la SOM Toolbox: el paquete básico y las funciones adicionales. El paquete básico está pensado para un uso independiente y dispone de documentación precisa y detallada. Excepto para las rutinas de menor nivel, cada función incluye, al comienzo del fichero, una pequeña descripción sobre sus tareas así como una ayuda más extensa inmediatamente a continuación [Vesanto et al. (2000)].

El tipo de información manejada por la herramienta SOM Toolbox son ficheros *.csv* o *.txt*. Cada fila de la tabla es una muestra de los datos de entrada. Los elementos en la fila son las variables que definen cada muestra. La relación entre los datos se refleja en la Figura [8]. Es importante que todas las muestras contengan el mismo número de variables. De este modo, cada columna alberga los valores de las variables que intervienen en el estudio. Algunos de estos valores pueden haberse perdido, pero la gran parte deberán incluirse en el set de datos.

La SOM Toolbox puede manejar tanto símbolos como datos numéricos, pero sólo este último tipo es utilizado en el algoritmo del SOM. Los datos denotados con símbolos pueden ser tratados como etiquetas asociadas a cada muestra de datos, aunque son ignoradas por la algorítmica del programa. Si se necesita trabajar con variables simbólicas en el entrenamiento del SOM, la mejor forma de tratar estos datos es convirtiéndolos en variables numéricas. Por ejemplo, sean los símbolos A, B y C. Si cada símbolo recibe como valores numéricos 1, 2 y 4 respectivamente, entonces B estará entre A y C y la distancia entre A y B será menor que la distancia entre B y C [Vesanto et al. (2000)].

Para el desarrollo de esta aplicación de han modificado algunas de las funciones base de la SOM Toolbox para conseguir un mejor funcionamiento del programa. Además, se han incluido nuevas tareas, implementadas en su totalidad, que facilitan la ejecución y permiten la división del código global en subtareas, lo que ayuda a comprender mejor el funcionamiento, así como a detectar de manera más fácil y rápida la aparición de errores en la ejecución.

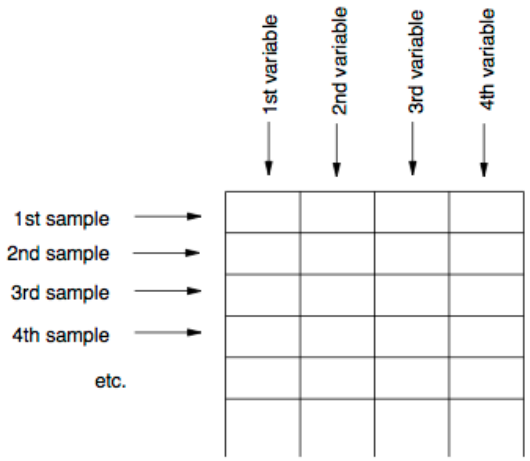


Figura [8] – Disposición de los datos de entrada en hojas de cálculo o tablas.

1.5. Discriminantes

Con la finalidad de estudiar las variables de entrada y establecer un orden de influencia sobre la conducta suicida, será necesario el uso de un método capaz de seleccionar, del conjunto de estudio, las variables más correlacionadas con el suicidio.

De acuerdo con la estadística clásica, una de las formas más simples de clasificación es el discriminante de Fisher. Para un problema de decisión binaria, el discriminante de Fisher de una variable v viene dado por la Ecuación [11] [Fukunaga (1990)].

$$d_F = \frac{|\mu_1(v) - \mu_0(v)|}{\sigma_1(v) + \sigma_0(v)} \qquad [11]$$

El criterio de Fisher es un discriminante lineal, es decir, es invariante a escalas. Por este motivo, es una aplicación limitada en problemas en los que ciertas variables están fuertemente relacionadas con la variable auxiliar si la relación es no lineal. Además, con este método de clasificación, los mapas autoorganizados no intervienen en la toma de decisiones.

Este hecho ha provocado el desarrollo en este proyecto de un segundo discriminante basado en el anterior. Este nuevo discriminante incorpora un término en la ecuación relativo a los centroides o celdas del mapa. A cada celda del mapa se asocia un vector $\hat{c}$ de las mismas dimensiones que los datos de entrada. El uso de estos vectores hará posible la intervención de los mapas autoorganizados en la selección de variables de interés. De hecho, se espera que el conjunto de variables destacadas sea diferente en ambos discriminantes para justificar así la aplicación de los SOM en la caracterización de la conducta suicida. En la Ecuación [12] se representa el discriminante de Fisher basado en SOM para una variable v.

$$d_{F_{SOM}} = \frac{|\hat{c}(v) - \mu_0(v)|}{\sigma_1(v) + \sigma_0(v)} \qquad [12]$$

Por otro lado, se ha desarrollado otro criterio propio de selección aún más centrado en los mapas autoorganizados. Este discriminante introduce una nueva expresión relativa a los histogramas de casos suicidas y no suicidas. El histograma $\hat{h}_1$ corresponde a la distribución de las muestras positivas, mientras que el parámetro $\hat{h}_0$ hace referencia a los casos negativos de los datos de estudio. Mediante la diferencia de ambos histogramas normalizados, se pretende detectar picos que permitan identificar perfiles suicidas o no suicidas. El producto de esta diferencia por el vector de centroides $\hat{c}(v)$ de cada variable ayudará a encontrar la correlación entre muestras que resulten de interés para el estudio. La intención es demostrar de manera gráfica la relación entre ciertas variables de estudio y la variable del comportamiento suicida. La Ecuación [13] define el discriminante basado en histogramas para una variable v.

$$d_{Hist}(v) = \frac{\sum_i \hat{c}_i(v)\left|\hat{h}_1^i - \hat{h}_0^i\right|}{\sigma(v)} \qquad [13]$$

1.6. Marco Regulador

La información resulta imprescindible para realizar con eficacia todas las tareas a las que se ha de dar respuesta a diario. El acopio de información presenta múltiples proyecciones y, en todas ellas, debe preservarse la privacidad de su contenido. Si la información versa sobre las personas, es decir, se refiere a datos de carácter personal ha de someterse a principios y reglas y controlarse para no provocar daños en los derechos de los individuos.

Si esa información incorpora además revelaciones sobre la salud, las garantías y cuidados deben extremarse. Los datos sobre la salud constituyen un elemento intrínseco y primordial en la vida de una persona. Con carácter general, la asistencia sanitaria prima en atención a la vida, pero deberá complementarse una norma con otra con el fin de equilibrar los intereses en juego y lograr la protección adecuada de todos los elementos afectados [Serrano (2005)].

1.6.1. Marco Legislativo Nacional e Internacional

La necesidad de armonizar la asistencia sanitaria y la protección de datos ha sido objeto de preocupación y atención. Desde un punto de vista general, el *Convenio 108 del Consejo de Europa para la Protección de las Personas con respecto al Tratamiento Automatizado de Datos de Carácter Personal* (ratificado por España el 27 de Enero de 1984), contemplaba una definición de datos sanitarios.

Por otra parte, la *Directiva 95/46, de 24 de Octubre relativa a la protección de las personas físicas y en lo que respecta al tratamiento de los datos personales y a la libre circulación de estos datos*[3], recoge la necesidad de extremar las garantías y la protección cuando la información se refiera a datos sobre la salud. Por último, la *Carta Europea de Derechos Fundamentales*, de 7 de Diciembre de 2000[4], eleva a la categoría de derecho fundamental el derecho a la protección de datos personales.

[3] D.O.L. núm. 281, de 23 de Noviembre de 1995.
[4] D.O.C.E. núm. 364, de 18 de Diciembre de 2000.

De una forma más sectorial, el *Convenio sobre Derechos Humanos y Biomedicina* de 1997[5], proclama en su artículo 10 el derecho de todos al respeto a la vida privada en el ámbito de la salud y el derecho a conocer cualquier información recogida al respecto. De una forma más detallada, el documento *Principios Éticos de la Sanidad de la Información*, de 30 de Julio de 1999, elaborado por el Grupo Europeo de Ética de la Ciencia y de las Nuevas Tecnologías, eleva a la Comisión Europea un informe en el que se relaciona el manejo de información personal con la prestación asistencial, siendo necesarias ambas para el beneficio de la persona. Este documento señala cuatro principios básicos en el tratamiento de datos sobre la salud:

1. Recogida de los datos, siempre que sea posible, del propio interesado, aunque también ha de contemplarse como supuesto normal en el ámbito sanitario la recogida de los datos por parte de los familiares del paciente.

2. Control de los datos sobre la salud por parte del propio interesado, es decir, la concreción del derecho fundamental a la protección de datos en el campo de la salud.

3. Derecho de oposición al uso de los datos personales siempre que la finalidad del uso no corresponda con la de la recogida.

4. Justificación de la utilización de los datos personales en la proyección social de la salud.

En el ámbito interno, además de la Ley Orgánica 15/1999 y de la Ley 41/2002, hay que referirse a la Ley General sanitaria donde se contienen los grandes principios fundamentales a respetar en el ámbito sanitario. En concreto, los artículos 10 y 11 en lo no derogado por la Ley 41/2002 hacen referencia a la dignidad de la persona, la libertad individual, el respeto al derecho a la intimidad del paciente y a la garantía de la confidencialidad respecto de la información personal utilizada.

En materia de protección de datos sanitarios habrá que prestar atención a la Ley Orgánica 15/1999, que regula las reglas generales de los tratamientos de datos, sin atender a especificidades, y a la Ley 41/2002 que, además de regular cuestiones puramente sanitarias, especifica para el campo sanitario la legislación general de protección de datos contenida en la Ley Orgánica. Ambas constituyen el marco normativo interno de los datos sobre la salud y su tratamiento [Serrano (2005)].

1.6.2. Principios de la Protección de Datos de la LOPD

1. El principio de la confidencialidad

Si hubiese que destacar un principio fundamental del tratamiento de los datos sobre la salud, probablemente se correspondería con el principio de la confidencialidad, que se basa en una confianza personal con el médico y en una confianza en sus posibilidades de tratamiento. Junto a ello, la confidencialidad se preserva limitando el acceso a los datos solamente a los profesionales que realizan el tratamiento médico y los ter-ceros legítimamente autorizados para ello y fijando las características del secreto médico.

[5] B.O.E. núm. 251, de 20 de Octubre de 2000.

2. El principio de adecuación y pertinencia

El artículo 4.1 de la LOPD exige que los datos sean adecuados, pertinentes y no excesivos en relación con las finalidades para las que se hayan recabado. La recogida y tratamiento de datos sanitarios persiguen una finalidad principal muy clara que es, según el artículo 16.1 de la Ley 41, "garantizar una asistencia adecuada al paciente". Esta finalidad determina, a su vez, la pertinencia y adecuación de los datos que recoja. Así lo señala el artículo 15 en su primer apartado, "la información que se considere trascendental para el conocimiento veraz y actualizado del estado de salud del paciente".

Con independencia de la valoración que realice el profesional médico respecto de la información trascendental para la asistencia sanitaria, la ley recoge un contenido mínimo referido a:

- La documentación referida a la hoja clínico-estadística.
- La autorización de ingreso.
- El informe de urgencia.
- La anamnesis[6] y la exploración física.
- La evolución.
- Las órdenes médicas.
- La hoja de interconsulta.
- Los informes de exploración complementaria.
- El consentimiento informado.
- El informe de anestesia.
- El informe de quirófano o de registro del parto.
- El informe de anatomía patológica.
- La evolución y planificación de cuidados de enfermería.
- La aplicación terapéutica de enfermería.

3. El principio de exactitud

El artículo 4.3 de la LOPD señala que los datos han de ser exactos y puestos al día de forma que respondan con veracidad a la situación real del interesado, obligación que en el terreno de la sanidad resulta especialmente importante en atención al interés vital que tiene el interesado en adecuar las informaciones incluidas en su historia clínica con su realidad.

4. El principio de cancelación

El artículo 4.5 de la LOPD recoge la cancelación de los datos cuando hayan dejado de ser necesarios o pertinentes en relación con la finalidad para la que se recogieron. La conexión entre el mantenimiento de los datos y su finalidad presenta en el ámbito sanitario alguna peculiaridad, pues es posible que sea preciso mantener los datos sanitarios en caso de tratamientos médicos prolongados o enfermedades crónicas en las que nunca llega a romperse la conexión entre la finalidad y el mantenimiento de los datos.

---

[6] La anamnesis es el térmico médico referido a la información proporcionada por el propio paciente al enfermero/a o médico durante la entrevista clínica. Comprende datos como la identificación, el motivo de consulta, la enfermedad actual, los antecedentes personales y familiares así como la revisión por sistema que consta de cuatro partes: inspección u observación, palpación, percusión y auscultación.

El artículo 17 de la Ley 41/2002 especifica la obligación de los centros sanitarios de conservar la documentación clínica en condiciones que garanticen su correcto mantenimiento y seguridad, aunque no necesariamente en soporte original, por el periodo necesario en cada caso y establece un periodo mínimo de 5 años desde el momento del alta del paciente.

5. **El principio de lealtad**

La información sanitaria, de acuerdo con el artículo 4.7 de la LOPD, no puede recogerse de forma desleal, fraudulenta o ilícita, criterio de legalidad que se extiende obviamente a la recogida de cualquier tipo de dato.

6. **El principio de seguridad**

El principio de seguridad es una garantía de la integridad de la información, de la disponibilidad de la misma y de su confidencialidad. Constituye un elemento importante en orden de preservar el derecho a la protección de datos personales. La LOPD regula en el artículo 9 el principio de seguridad. Por su parte la Ley 41/2002 realiza una completa regulación de la seguridad de las historias clínicas recogiéndola como un deber para los centros médicos [Serrano (2005)].

1.7. Entorno Socioeconómico

El 10 de Septiembre, Día Mundial para la Prevención del Suicidio, se fomentan en todo el mundo compromisos y medidas prácticas para prevenir los suicidios. Cada día hay, en promedio, casi 3.000 personas que ponen fin a su vida y, al menos, 20 personas intentan suicidarse por cada una que lo consigue [WHO (2012)].

En los últimos 45 años, las tasas de suicidio han aumentado en un 60% a nivel mundial. El suicidio es una de las tres primeras causas de defunción entre las personas de 15 a 44 años en algunos países, y la segunda causa en el grupo de 10 a 24 años, sin contar con las tentativas de suicidio, que son hasta 20 veces más frecuentes que los casos de suicidio consumado [WHO (2012)].

Se estima que a nivel mundial el suicidio supuso el 1,8% de la carga global de morbilidad en 1998, y que en 2020 representará el 2,4% en los países con economías de mercado y en los antiguos países socialistas [WHO (2012)].

Aunque tradicionalmente las mayores tasas de suicidio se habían registrado entre los varones de edad avanzada, las tasas entre los jóvenes han ido en aumento hasta el punto que ahora ellos constituyen el grupo de mayor riesgo en un tercio de los países, tanto en el mundo desarrollado como en el mundo en desarrollo [WHO (2012)].

Los ancianos constituyen uno de los principales grupos de riesgo por varios factores, como la pérdida de poder adquisitivo que va aparejada a la jubilación o una esperanza de vida cada vez mayor, que desencadena enfermedades crónicas, problemas familiares o la pérdida del cónyuge. En el caso de los adolescentes, el fracaso escolar, un desengaño amoroso, el divorcio de los padres o conductas de imitación pueden convertirse en desencadenantes de la muerte voluntaria [Mengual (2011)].

Tasas de suicidio en distintos países

Suicidios por 100.000 personas.

	Hombres	Mujeres	Total
Korea	49,6	21,4	33,5
Hungría	40,4	9,5	23,3
Rusia	42,1	7,2	22,4
Japón	31,4	11,5	21,2
Eslovenia	32,8	6,6	18,6
Bélgica	26,6	9,5	17,7
Finlandia	26,8	8,3	17,3
Suiza	24,8	10,2	16,9
Francia	25,7	8,0	16,2
Polonia	29,2	4,0	15,9
Estonia	29,3	5,0	15,8
Australia	23,3	6,0	13,9
Rep. Checa	23,8	4,4	13,5
Chile	22,4	5,0	13,3
OECD	**20,8**	**5,9**	**12,9**
Nueva Zelanda	18,9	6,4	12,4
Estados Unidos	19,8	4,9	12,0
Islandia	18,7	4,6	11,8
Suecia	17,5	6,1	11,7
Dinamarca	18,0	6,1	11,6
Luxemburgo	17,6	5,2	11,3
Eslovenia	20,7	3,2	11,3
Noruega	15,9	6,7	11,2
Cananda	17,3	5,1	11,1
Irlanda	17,4	4,6	11,0
Alemania	17,3	5,1	10,8
Austria	16,7	4,8	10,6
Portugal	15,8	4,2	9,3
Países Bajos	13,3	5,4	9,2
Reino Unido	10,5	3,0	6,7
España	**10,5**	**2,7**	**6,3**
Israel	9,9	2,8	6,2
Italia	9,8	2,5	5,9
Brasil	9,0	2,3	5,4
Mexico	8,5	1,5	4,8
Grecia	5,6	0,9	3,2

Figura [9] – Tasas de suicidio en distintos países en el 2010 (OECD).

Por otro lado, es importante los datos de tasas de suicidio en mujeres y hombres. Las mujeres lo intentan más, pero los hombres son más efectivos y utilizan métodos más contundentes. La tasa general de suicidio entre las mujeres no ha cambiado en los últimos años, manteniéndose en el 1,3% del total de las muertes, mientras que el número de suicidios entre los hombres ha aumentado significativamente del 2,7% hasta el 3,9% en 2001 y el 5,8% en 2010 [Mengual (2011)] [Associated Press (2013)].

Los trastornos mentales, especialmente la depresión y los trastornos por consumo de alcohol, son un importante factor de riesgo de suicidio en Europa y América del Norte. En los países asiáticos, sin embargo, tiene especial importancia la conducta impulsiva. El suicidio es un problema complejo en el que intervienen factores psicológicos, sociales, biológicos, culturales y ambientales [WHO (2012)].

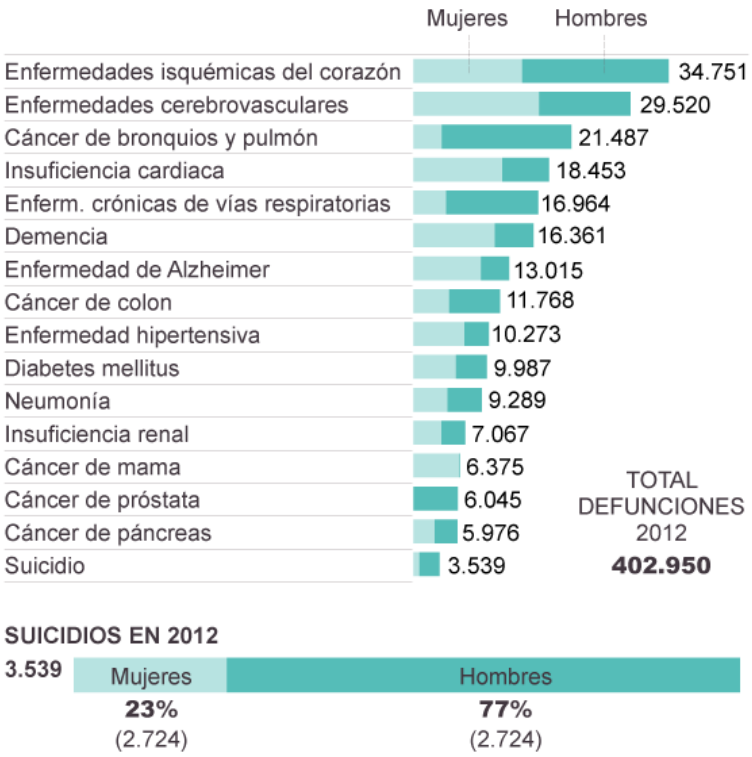


Figura [10] – Causas de muerte más frecuentes en hombres y mujeres durante el 2012.

En el 90-95% de los casos existe algún tipo de trastorno psiquiátrico, la mayor parte de las veces, una depresión. De ahí la importancia de la atención y detección temprana. El 5% restante obedece a un factor existencial que hace que la persona en cuestión vea en el suicidio la única manera de poner fin a sus problemas. Los antecedentes familiares, padecer una enfermedad crónica con dolor, conductas adictivas, acontecimientos vitales que suponen pérdidas afectivas, el aislamiento y el hecho de haber tenido alguna vez pensamientos suicidas son otros factores de riesgo [Mengual (2011)].

Científicos del Centro de Investigación y Prevención del Suicidio, de la Universidad de Hong Kong, concluyeron un estudio en el cual se demuestra que la crisis económica desencadenada en 2008 ha provocado el aumento de suicidios a nivel mundial. Los resultados son preocupantes, pues los desastres económicos han elevado significativamente el listón del número de suicidios a escala global [Associated Press (2013)].

A nivel mundial, la prevención del suicidio es una necesidad que no se ha abordado de forma adecuada debido básicamente a la falta de sensibilización sobre la importancia de ese problema y al tabú que lo rodea e impide que se hable abiertamente de ello. De hecho, solo unos cuantos países han incluido la prevención del suicidio entre sus prioridades [WHO (2012)].

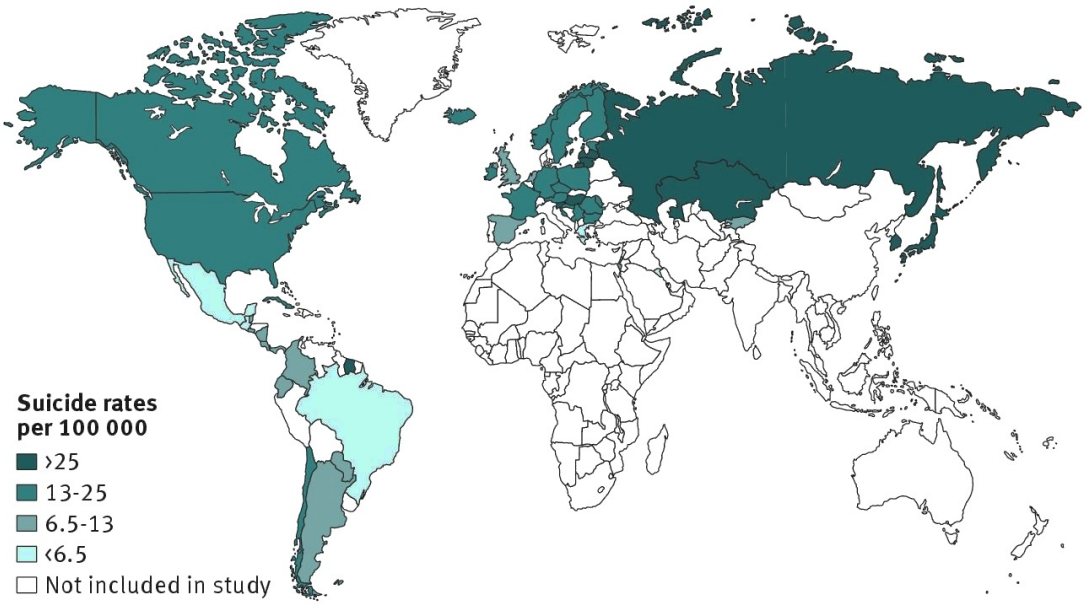


Figura [11] – Mapa con los 54 países analizados por el Centro de Investigación y Prevención del Suicidio de la Universidad de Hong Kong y su tasa de suicidios en 2009.

Por eso, la OMS, la ONU y la Unión Europea han lanzado la voz de alerta y señalado la muerte voluntaria como un problema de salud pública de primera magnitud. La OMS demanda que autoridades y gobiernos adopten medidas de prevención, dado que las cifras demuestran que las actuales son insuficientes [Mengual (2011)].

Es evidente que la prevención del suicidio requiere también la intervención de sectores distintos del de la salud y exige un enfoque innovador, integral y multisectorial, con la participación tanto del sector de la salud como de otros sectores tales como la educación, el mundo laboral, la policía, la justicia, la religión, el derecho, la política y los medios de comunicación [WHO (2012)].

A nivel nacional, el número de fallecidos por suicidio aumentó un 11,3% durante 2012, según los datos publicados por el Instituto Nacional de Estadística. La subida no es excesiva pero sí relevante si se tiene en cuenta que, en el año anterior, el incremento fue sólo del 0,7% [San-Martín (2014)].

De acuerdo a la cifras de defunciones según la causa de muerte del INE, un total de 3.529 personas murió por suicidio en 2012, 2.724 hombres y 815 mujeres. La tasa se sitúa en 7,6 por cada 100.000 personas. El suicidio es la principal causa externa de mortalidad desde hace unos años, cuando las muertes por accidentes de tráfico empezaron a descender [San-Martín (2014)].

Se ha alertado desde distintos ámbitos de que el número de suicidios podía estar aumentando por la crisis económica, pero la realidad es que en 2012 es cuando aparece por primera vez esta tendencia. En 2011 el incremento fue muy pequeño y en 2010 se registró la cifra más baja en casi dos décadas [San-Martín (2014)].

En su informe, el INE señala que la tasa bruta de mortalidad en España en 2012 ha sido un 3,8% superior con respecto al año anterior. Las muertes que más se han incrementado han sido por trastornos mentales y del comportamiento, como demencias vasculares, demencias seniles y otras distintas al Alzheimer, y las enfermedades del sistema respiratorio, ambas en un 12% [San-Martín (2014)].

La provincia que registra una mayor tasa de suicidios en 2012 es Lugo, seguida de Granada, Galicia y Asturias. Las provincias del centro de España, con Madrid a la cabeza, son las que registran una menor tasa de suicidios [Massot (2014)].

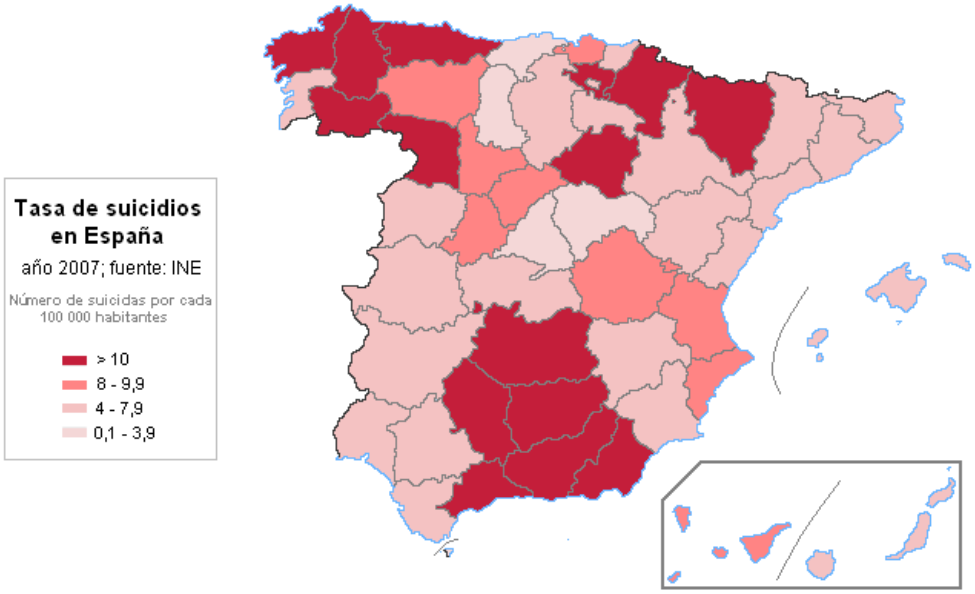


Figura [12] – Tasa de suicidios en España en 2007 (INE).

1.8. Objetivos del Proyecto

Una vez analizada la necesidad por la que surge esta aplicación, así como la metodología a seguir para el desarrollo y consecución de la misma, se plantean una serie de objetivos que deberán cumplirse al término del proyecto.

- ✓ Caracterización del comportamiento suicida mediante Mapas Autoorganizados:

  Tras comprobar la validez de los Mapas Autoorganizados para el estudio de datos de grandes dimensiones con relaciones fuertemente no lineales, se pretende visualizar gráficamente las variables de mayor interés que caractericen la conducta suicida, identificando perfiles de sujetos que permitan configurar un modelo generalizado de individuo potencialmente suicida.

- ✓ Desarrollo de una herramienta de análisis y visualización:

  La principal tarea de este proyecto comprende el desarrollo de una herramienta informática capaz de analizar y estudiar los datos de entrada, obteniendo como resultado una serie de gráficas y parámetros que permitan visualizar la correlación entre variables de interés y la conducta suicida.

- ✓ Obtención de resultados de ayuda a la detección de intentos de suicidio:

  La finalidad de este proyecto es conseguir una relación de variables involucradas en el comportamiento suicida que ayuden a detectar y prevenir los intentos de suicidio. Los resultados obtenidos deberán ser evaluados desde un punto de vista médico, es decir, la información extraída de este programa no es 100% fiable, por lo que también es necesario el estudio de factores de riesgo y protección por personal médico cualificado. Si bien, se pretende aportar una herramienta que sea de utilidad en psiquiatría y facilite la tarea de diagnóstico en pacientes con alto riesgo de suicidio.

2. Descripción de la Solución

En los siguientes apartados se presentará el proceso de desarrollo de la solución dada al problema propuesto. En el apartado 2.1. se explicará cómo se ha realizado el preprocesado de datos y su adecuación a las características del código. En la sección 2.2. se describirá la interfaz de usuario implementada para la ejecución del programa. El apartado 2.3. recoge la metodología de inicialización y entrenamiento de los mapas. En el apartado 2.4. se especificará cuáles son los tipos de inicialización desarrollados, mientras que en la sección 2.5. se explicará la influencia del número de celdas del mapa. Por último, los apartados 2.6. y 2.7. contienen los criterios de selección de parámetros y variables relevantes para el estudio.

2.1. Datos de Estudio

Para el estudio de la conducta suicida, se dispone de un conjunto de 8.699 sujetos, cada uno de ellos caracterizado por 606 variables, entre las que se encuentra el género, la edad, el estado civil, el nivel de estudios, trastornos mentales, alcoholismo, impulsividad, traumas de infancia, etc. Para que la algorítmica del SOM sea capaz de trabajar con estos datos, se han dispuesto en una tabla como la que aparece en la Figura [8]. En la matriz de datos, cada columna hace referencia a una de las 606 variables, mientras que cada fila incluye una muestra de estudio, quedando definido un perfil para cada sujeto.

Dentro del conjunto de variables existen registros categóricos que han sido identificados numéricamente, lo que carece de sentido matemático. Para evitar problemas en la ejecución del programa, se ha creído oportuno realizar la dicotomización de algunas de estas variables de manera que la interpretación de las mismas sea lo más correcta posible. Es el caso de la variable EST_CIV, que define el estado civil del sujeto, siendo 1 si está soltero, 2 si está casado, 3 si está divorciado y 4 si es viudo. El cambio efectuado ha supuesto la creación de cuatro variables nuevas: EST_CIVIL_1, EST_CIV_2, EST_CIV_3 y EST_CIV_4, que indican con un 1 si el sujeto está soltero, casado, divorciado o viudo, respectivamente, y con un 0 si no cumple con ese estado civil. La variable EST_CIV ha sido suprimida de la base de datos.

De este modo, se dispone en total de un conjunto de 609 variables. Cada una de estas variables será contrastada con la variable suicida, que identifica con 1 los intentos de suicidio y con 0 el caso contrario. Por tanto, considerando también el factor suicida, el conjunto de datos de estudio forma una matriz de 8.699 filas y 610 columnas.

Además, se define un vector de etiquetas en el que se incluyen los nombres de las variables de estudio. El algoritmo del SOM no utiliza estas etiquetas en sus cálculos, pero serán de ayuda en la identificación de las variables de interés y visualización de gráficas.

Los datos de entrada no pueden tratarse en bruto, si no que precisan de una normalización, de modo que todos los valores oscilen entre 0 y 1 para evitar problemas de escalas. Para normalizar las muestras de estudio, la herramienta SOM Toolbox incorpora una función denominada `som_normalize.m`, que es la que se ha utilizado en el desarrollo de este proyecto para la normalización de lo datos, con los que previamente se ha construido una estructura mediante el comando `som_data_struct.m`. La estructura de datos incluye diferentes campos que la algorítmica del SOM necesita para llevar a cabo la ejecución del código.

Por otro lado, algunos de los valores de las variables de estudio se corresponden con datos perdidos, lo que en el lenguaje matemático de Matlab se identifica como NaN (*Not a Number*). Estos valores indefinidos producen problemas en algunos procesos del código, de modo que, dependiendo de los puntos del programa en los que los datos perdidos generan errores en la ejecución, se ha optado por omitirlos o sustituirlos por la media correspondiente a cada variable. Es decir, para cada variable de estudio se determina la media prescindiendo de los valores perdidos y, posteriormente, se reemplazan estos valores por la media calculada. El proceso de sustitución de valores indefinidos se denomina imputación de datos. De esta forma, se evitan problemas en la compilación y ejecución del código.

El listado de variables cuenta con la aprobación del consorcio EURECA, en el que participan cinco instituciones en la investigación para la detección y prevención de la conducta suicida.

2.2. Interfaz de Usuario

Ante las múltiples posibilidades de ejecución en las que se ha trabajado, se ha considerado importante la inclusión de un menú de selección de parámetros que se muestra por línea de comandos cada vez que se ejecuta el programa. En la Figura [13] se representa una captura de la pantalla de ejecución del código en la plataforma Matlab.

Al comienzo, se muestra una breve descripción de los Mapas Autoorganizados así como de las funciones que realiza el programa. Para que el ejecutor de la aplicación comprenda la finalidad del experimento, se ha fijado un tiempo de pausa para leer la introducción al programa. Una vez finalice la lectura, se podrá continuar con la ejecución presionando cualquier tecla.

A continuación, se inicia la selección de parámetros. Es interesante especificar el tamaño del SOM, así como el tipo de inicialización. En cuanto a las dimensiones, se han definido hasta cinco posibles combinaciones de celdas, siguiendo por defecto un patrón hexagonal. Por otro lado, existen tres tipos de inicializaciones posibles, que serán analizadas en el siguiente apartado.

Para que el código sea robusto y corregir errores en los parámetros de entrada por teclado, el programa ha sido diseñado de modo que, en caso de fallo, el ejecutor tenga una nueva oportunidad en la entrada de caracteres. De este modo, si las opciones disponibles en el caso de las dimensiones son 1, 2, 3, 4 ó 5, y el usuario introduce un número mayor o menor, un símbolo o una cadena de caracteres y, a continuación pulsa *enter*, el programa le permitirá teclear de nuevo el parámetro demandado, tantas veces como errores cometa el ejecutor.

```
                    ::: MAPAS AUTOORGANIZADOS :::

Los Mapas Autoorganizados (SOM) son un tipo de red neuronal
no supervisada que consiste en una cuadricula bidimensional
de celdas, cada una caracterizada por un vector de la misma
dimensión que los datos.

La formación de la red se realiza iterativamente utilizando
los datos de entrenamiento en dos etapas distintas: cálculo
del Best Matching Unit (BMU) y actualización de centroides.

El programa evalúa un conjunto de datos para determinar las
variables mas influyentes en el comportamiento suicida. Las
muestras evaluadas incluyen, además de intentos de suicidio,
pacientes psiquiátricos sin antecedentes de intentos suici-
das, sin trastornos mentales o conductas suicidas, donantes
de sangre y familiares de personas con intentos de suicidio.

Pulse una tecla para continuar...

PARAMETROS DE ENTRADA:

* Dimensiones del SOM

      1. 6x2
      2. 10x6
      3. 16x12
      4. 24x20
      5. 50x46

      Opción: -1

      Error. El carácter introducido no es valido.

      Opción: 0

      Error. El carácter introducido no es valido.

      Opción: 6

      Error. El carácter introducido no es valido.

      Opción: a

      Error. El carácter introducido no es valido.

      Opción: hola

      Error. El carácter introducido no es valido.

      Opción: 3


* Tipo de Inicialización

      1. Aleatoria
      2. Lineal
      3. Proyección LDA

      Opción: 1
```

Figura [13] – Pantalla de ejecución del programa en Matlab.

2.3. Inicialización y Entrenamiento

Partiendo de una matriz de datos de dimensiones 8699x609 y de un vector contenedor de la variable suicida definido para 8699 perfiles, la adecuación de estas variables pasa por configurar una matriz común en la que se agrupan todos los factores que intervienen en el estudio, incluida la variable que representa el comportamiento suicida, tal y como se representa en la Figura [14].

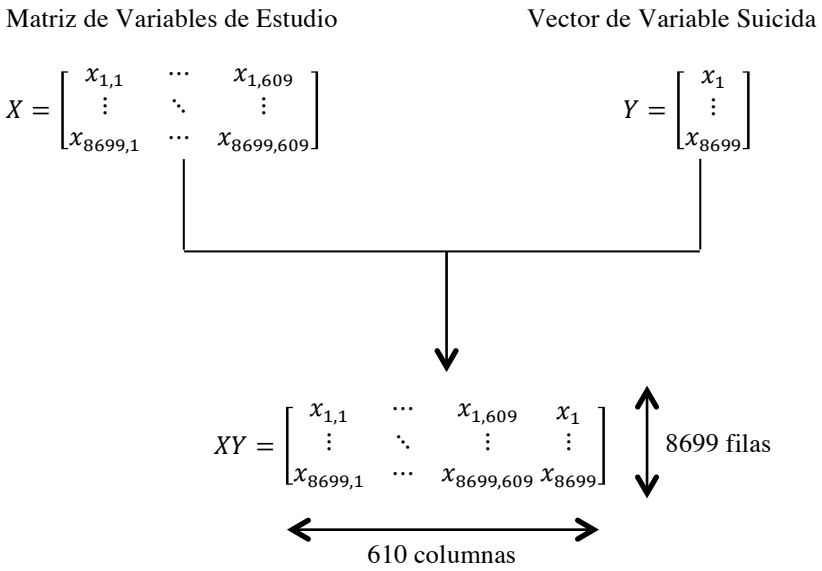


Figura [14] – Disposición de los datos de estudio en el código.

Una vez dispuestas todas las variables en una única matriz, se conformará una estructura de datos con la que el algoritmo SOM pueda trabajar. Además, las variables se normalizarán utilizando la función proporcionada por la herramienta de trabajo SOM Toolbox para que todos los valores oscilen entre 0 y 1.

El siguiente paso será el enmascaramiento de la variable suicida para que no intervenga en las fases de inicialización y entrenamiento y, por tanto, no influya en la agrupación de las celdas del mapa. La variable *Y* se incluye en los datos de estudio únicamente para comparar los factores de interés con la conducta suicida. Mediante el uso de esta variable, se podrán diferenciar los perfiles de la base de datos en función de si definen un comportamiento suicida o no. Es decir, fila a fila, se irá comprobando si la variable suicida toma valor 0 o 1. Si tiene valor 1, el perfil se clasificará como suicida, mientras que si el valor de la variable *Y* es 0, el perfil se identificará como una conducta no suicida. De esta forma, se conformarán dos subgrupos, uno de muestras positivas y otro de muestras negativas. Además, mediante las gráficas de la variable suicida y del resto de factores, podrán observarse las similitudes o diferencias entre zonas frías y calientes del mapa.

El proceso de inicialización consiste en otorgar a cada celda un vector de centroides de las mismas dimensiones que los datos de entrada. Para inicializar cada una de las celdas, se han definido tres métodos de inicialización, que serán descritos en el siguiente apartado.

Por otro lado, la fase de entrenamiento se encargará de la búsqueda del BMU y de la actualización de centroides, o lo que es lo mismo, de la agrupación de celdas en función de las similitudes entre sus vectores.

Así, para cada variable se definirá un valor de centroide en cada celda del mapa. Cuanto mayor sea el valor del centroide, el color de la celda será de un rojo más intenso y, por tanto, mayor relevancia tendrá dicha variable de estudio. Por otro lado, valores menores de centroides se identificarán con colores verdes y azules, simbolizando de esta forma la no influencia de esos factores sobre la conducta suicida.

En el siguiente mapa se representa la variable suicida con un método de inicialización lineal y unas dimensiones de 16x12 celdas. Puede observarse la variedad de colores en el mapa y la identificación de zonas de interés o puntos calientes.

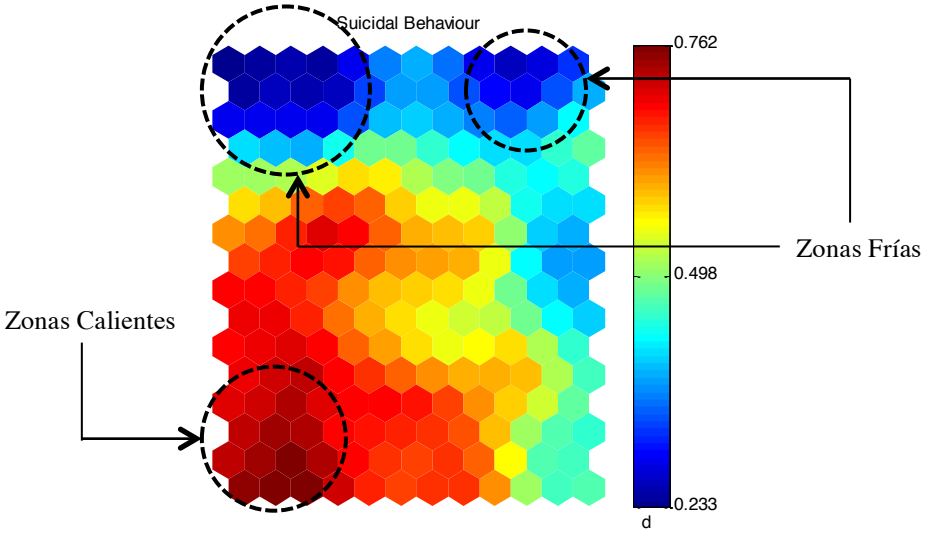


Figura [15] – Mapa autoorganizado de la variable suicida.

2.4. Tipos de Inicialización

El primer paso en la creación de un SOM comprende la inicialización de sus celdas. Este proceso consiste en asociar a cada celda un vector de las mismas dimensiones que los datos de entrada. Una vez que todas las celdas del mapa estén caracterizadas por un vector $\boldsymbol{m}_i$, se iniciará la búsqueda del BMU y se actualizarán los valores de los centroides, situándose más próximas las celdas con vectores entre los que exista una mayor similitud. Es decir, dado un vector de entrada, se calculará su distancia Euclídea con respecto a los vectores asociados a cada celda. Una vez se haya localizado el *Best Matching Unit*, el centroide correspondiente se desplazará hacia zonas del mapa en las que los vectores modelo sean muy parecidos, conformándose así vecindarios de gran similitud.

La finalidad de definir diferentes tipos de inicialización pasa por tratar de conseguir unos valores iniciales bastante parecidos a los valores que quieren obtenerse tras el entrenamiento del mapa. De este modo, se estarían suprimiendo iteraciones de código con la consiguiente reducción del tiempo de ejecución.

Los tres tipos de inicialización con los que se ha trabajado son aleatoria, lineal y con proyección LDA. Los dos primeros son modelos implementados por la herramienta SOM Toolbox. Si bien, se han efectuado ciertos cambios en el código que define el criterio lineal, con la finalidad de incorporar la proyección LDA en la misma función y economizar las tareas ejecutadas por el programa.

2.4.1. Inicialización Aleatoria

El método de inicialización aleatoria está implementado en la función `som_randinit.m` incluida en el paquete de la SOM Toolbox de Matlab. Recibe como parámetros los datos de entrada así como la estructura del mapa, y su tarea consiste en inicializar el SOM con valores aleatorios. Para cada componente x_i, los valores de los vectores asociados a cada celda del mapa se distribuyen uniformemente en el rango $[\min(x_i),\ \max(x_i)]$ [Vesanto et al. (2000)].

2.4.2. Inicialización Lineal

La inicialización lineal del mapa se implementa en la función `som_lininit.m` también incluida en la herramienta SOM Toolbox. En su versión por defecto, recibe como parámetros los datos de entrada y el mapa a inicializar, al igual que en el modelo aleatorio. La inicialización se efectúa calculando en primer lugar los autovectores y autovalores de los datos de entrada.

Siendo A el vector definido por los datos de estudio, un vector propio de A es un vector $\boldsymbol{x}$ distinto de cero tal que $A\boldsymbol{x} = \lambda\boldsymbol{x}$ para algún $\lambda \in \mathbb{R}$. A λ se le llama valor propio asociado a A. El método lineal inicializa los vectores asociados a las celdas del mapa a lo largo de los $mdim$ autovectores con mayores autovalores. El parámetro $mdim$ coincide con las dimensiones del mapa, siendo 2 el valor más usual [Vesanto et al. (2000)].

La inicialización lineal realiza, por tanto, un barrido desde la esquina superior izquierda del mapa tal y como se muestra en la Figura [16]. Los autovectores van recorriendo las distintas celdas del mapa e inicializando los modelos $\boldsymbol{m}_i$ asociados a cada centroide.

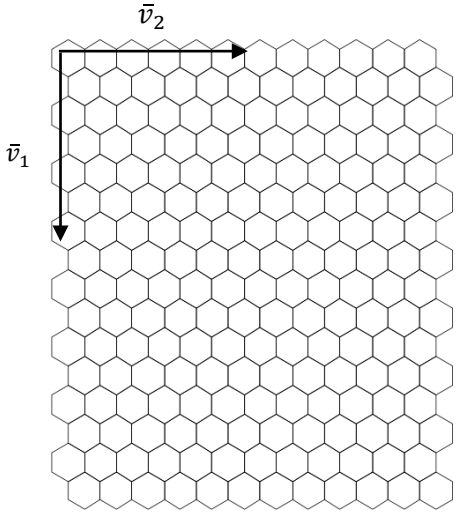


Figura [16] – Inicialización de celdas con el método lineal.

2.4.3. Inicialización con Proyección LDA

De acuerdo a las necesidades del proyecto, se ha diseñado totalmente un nuevo criterio de inicialización. Basado en el modelo lineal, el criterio de inicialización con proyección LDA pretende aplicar el mismo principio que el método anterior pero utilizando en este caso vectores más relacionados con las muestras de estudio.

La proyección LDA (*Linear Discriminant Analysis*) es un conocido esquema utilizado en la extracción de características así como en la reducción de dimensiones. Este método ha sido ampliamente utilizado en numerosas aplicaciones como reconocimiento facial o recuperación de imágenes. El modelo de proyección LDA clásico proyecta los datos sobre un vector espacial de reducidas dimensiones de modo que la relación entre la distancia entre clases y las distancias dentro de las clases se maximiza, logrando así la discriminación máxima [Jieping et al. (2004)].

Para el cálculo de la proyección LDA se ha diseñado una nueva función denominada `som_LDA.m`, que recibe como parámetros la matriz de datos de entrada sin incluir la variable suicida, el vector que almacena los datos de la conducta suicida y el número de componentes a extraer. Como resultado de la ejecución de esta función, se obtendrá el vector de proyección LDA.

A continuación, por medio del código que implementa el criterio lineal, se aplicará este nuevo modelo de inicialización. Se modifican, por tanto, los parámetros de entrada de la función `som_lininit.m`. La nueva estructura incluye, además de los datos de estudio y la referencia al mapa, un parámetro que identifica si se está llamando a dicha función para una inicialización lineal (`W=0`) o una inicialización con proyección LDA (`W=1`), así como el vector de la proyección obtenido mediante la función `som_LDA.m`, siendo nulo en el caso de inicialización lineal.

La finalidad con la que se implementa este nuevo método consiste en sustituir uno de los autovectores calculados en inicialización lineal mediante el criterio de vectores y valores propios. El vector de proyección LDA ($\overline{w}_{LDA}$) pasa a ocupar la posición de la segunda componente lineal, mientras que el autovector con mayor autovalor ($\bar{v}_1$) sigue manteniéndose en la misma dirección que en el modelo lineal. Con estos cambios, la nueva inicialización del mapa seguirá un patrón como el que se muestra en la Figura [17].

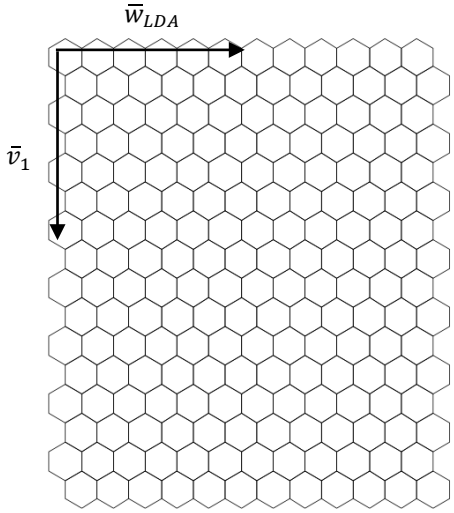


Figura [17] – Inicialización de celdas con el método de proyección LDA.

2.5. Número de Celdas del SOM

Un parámetro a tener en cuenta en los Mapas Autoorganizados es el tamaño o número de celdas que lo integran. Se han considerado distintas dimensiones con el objetivo de encontrar el tamaño óptimo de mapa que proporcionase resultados similares a los esperados. Un número de celdas pequeño apenas aporta información relevante para el estudio. Un mapa de grandes dimensiones tiene mayor resolución pero, sin embargo, proporciona demasiados datos que enmascaran ciertas características de las variables de entrada.

Por lo tanto, un tamaño óptimo de mapa es aquél de dimensiones intermedias. Para el estudio de este proyecto se ha optado por un tamaño de 16x12 celdas [filas x columnas], aunque se han realizado experimentos para distintas dimensiones. Los criterios desarrollados para valorar la validez de las combinaciones de celdas implementadas se detallan en el siguiente apartado.

2.6. Criterios de Selección de Parámetros

Dados los diferentes tipos de inicialización y tamaños de mapa, se precisa el desarrollo de criterios que faciliten la toma de decisiones y la selección de los parámetros óptimos para la realización del estudio. En base a estas necesidades, se han implementado dos métodos de decisión basados en las diferencias existentes entre los histogramas de casos positivos, es decir, de comportamiento suicida, y negativos, o lo que es lo mismo, individuos sin presencia de intentos de suicidio.

2.6.1. Criterio del Coseno

Este criterio está basado en la distancia entre los histogramas de muestras positivas y negativas. Siendo $\bar{h}_1$ el histograma que define los casos de intentos de suicidio y $\bar{h}_0$ el correspondiente a las muestras sin presencia de conducta suicida, la distancia Euclídea entre ambos vectores se define en la Ecuación [14].

$$d^2(h_1,h_0) = \left\|\bar{h}_1 - \bar{h}_0\right\|^2 = \left\|\bar{h}_1\right\|^2 + \left\|\bar{h}_0\right\|^2 - 2\left\langle \bar{h}_1, \bar{h}_0 \right\rangle \qquad [14]$$

Desarrollando el producto escalar entre los histogramas $\bar{h}_1$ y $\bar{h}_0$, la distancia es igual al producto de las normas de los vectores y el coseno del ángulo que forman, tal y como aparece en la Ecuación [15].

$$d^2(h_1,h_0) = \left\|\bar{h}_1 - \bar{h}_0\right\|^2 = \left\|\bar{h}_1\right\|^2 + \left\|\bar{h}_0\right\|^2 - 2\left|\bar{h}_1\right|\left|\bar{h}_0\right|\cos\alpha \qquad [15]$$

Si se igualan ambas expresiones y se despeja el coseno del ángulo α, se obtiene la expresión que se ha utilizado para el cálculo de este criterio, representada en la Ecuación [16].

$$\cos\alpha = \frac{\left\langle \bar{h}_1, \bar{h}_0 \right\rangle}{\left|\bar{h}_1\right|\left|\bar{h}_0\right|} \qquad [16]$$

Bajo este parámetro puede determinarse la similitud o diferencia entre ambos histogramas. Si son muy parecidos, entonces los vectores que representan serán paralelos y el ángulo que formen de 0º o 180º, luego el coseno tendrá valor 1 o -1 respectivamente. Por otro lado, si los histogramas son muy diferentes, es decir, el programa está discriminando correctamente entre casos positivos y negativos, los vectores serán perpendiculares, formando un ángulo de 90º, y el valor del coseno será 0. Por tanto, cuanto menor sea el valor del coseno del ángulo α mejor será la distinción que el código está haciendo entre perfiles suicidas y no suicidas.

Para el cálculo de este parámetro se ha implementado una nueva función en el código denominada `som_cosine_alpha.m`, que recibe como parámetros los vectores referentes a cada uno de los histogramas. A su salida devuelve el valor del coseno del ángulo.

2.6.2. Criterio de la Distancia

Este criterio está basado en la distancia de norma 1 y cuadrática entre los histogramas de muestras positivas y negativas, de modo que, cuanto mayor sea la distancia calculada, más diferentes serán ambos histogramas y mejor será la diferenciación realizada por el programa para la clasificación de sujetos con intentos de suicidio y sin presencia de conducta suicida.

Para determinar la distancia de norma 1 se ha implementado la expresión de la Ecuación [17]. Debido a que el cociente entre la norma de la diferencia de histogramas y el número de elementos coincide con la media, se ha aplicado esta simplificación al código del programa.

$$d_1 = \frac{\left|\hat{h}_1 - \hat{h}_0\right|}{N} \qquad [17]$$

Los histogramas $\hat{h}_1$ y $\hat{h}_0$ están normalizados por su norma, de modo que todos los valores oscilen entre 0 y 1. La distancia cuadrática se calcula aplicando la fórmula de la Ecuación [18].

$$d_2 = \frac{\left|\hat{h}_1 - \hat{h}_0\right|^2}{N} \qquad [18]$$

Para la implementación de este criterio se ha incluido una nueva función al paquete de la SOM Toolbox llamada `som_distance.m`. Esta función recibe como parámetros los histogramas de casos positivos y negativos sin normalizar así como el número de filas y columnas del mapa. El primer paso consiste en normalizar los histogramas dividiéndolos por su norma. A continuación, se determinan las distancias aplicando las expresiones de las Ecuaciones [17] y [18], donde el número de elementos se calcula como el producto de filas y columnas.

2.7. Criterios de Selección de Variables

El objetivo principal de este proyecto es elaborar una lista de aquellas variables más relevantes en el comportamiento suicida. Para determinar el orden de importancia de las variables de entrada, existen dos métodos que ayudarán a detectar los factores de riesgo en la conducta suicida.

2.7.1. Puntos Calientes

Mediante la observación de los mapas autoorganizados de cada una de las variables de estudio, pueden apreciarse zonas de mayor o menor concentración de muestras. Los puntos calientes son las zonas de mayor riesgo de suicidio y están representadas con tonalidades intensas de rojos. De este modo, será una tarea fácil localizar los factores más influyentes en la conducta suicida, ya que la celda con mayor intensidad de rojo, será la que contenga el centroide de mayor valor.

Por ejemplo, dada la variable `STAI_anxiety`, que representa la ansiedad del sujeto, en la Figura [18] puede observarse la similitud con respecto a la variable de la conducta suicida. En este caso, el parecido entre ambas gráficas tiene sentido, ya que una y otra variable se encuentran correlacionadas entre sí. A priori, sin la realización de estudios psiquiátricos, se asume que, cuanto mayor sea la ansiedad de una persona, mayor serán las probabilidades de un final suicida. Por este motivo, las zonas frías y los puntos calientes se solapan en los dos casos.

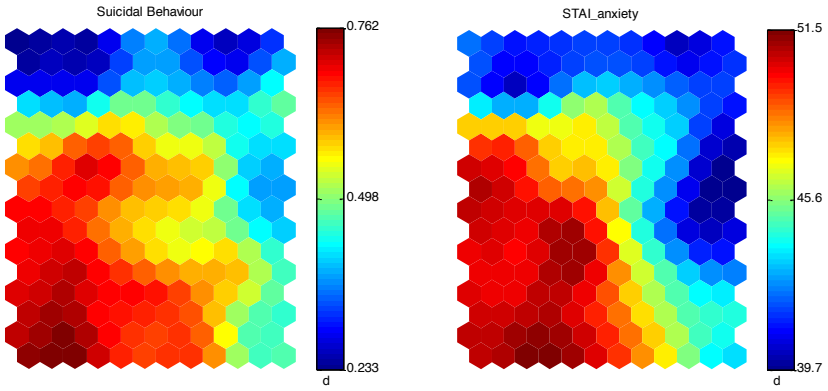


Figura [18] – Correlación entre zonas frías y puntos calientes en los SOM.

Si bien, algunas de las variables de estudio están extraídas de cuestionarios en los que las preguntas pueden estar formuladas en positivo o negativo. Estando formuladas en positivo, se obtendrá una correspondencia como en el caso anterior, siempre y cuando, la variable en cuestión resulte de interés. Sin embargo, se descubren nuevas variables cuyas gráficas resultan ser inversas a la representación del comportamiento suicida. En algunos de estos casos, los factores no serán de interés para el estudio, es decir, las variables no serán influyentes en la conducta suicida y, de ahí, sus diferencias gráficas.

No obstante, para determinadas variables, esta inversión en los mapas autoorganizados se interpretará como un planteamiento negativo en las preguntas de los correspondientes cuestionarios. Esto es, sea la variable `STAXI_Angecontrol` la responsable de caracterizar el control de enfados.

Analizando los resultados de manera lógica, si un sujeto no es capaz de contener su enfado, será una persona violenta e impulsiva, lo que supone un alto riesgo de suicidio. Sin embargo, observando las gráficas de la Figura [19] se comprueba que la zona caliente de la variable suicida, es decir, la de mayor riesgo de suicidio, se solapa con una zona fría del factor `STAXI_Angecontrol`. La razón lógica es que, cuanto mayor control tiene un sujeto del enfado, menores son las probabilidades de que esa actitud desencadene en un intento de suicidio.

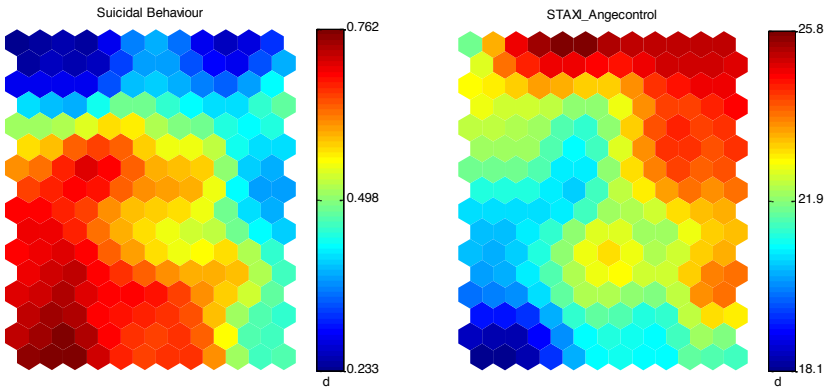


Figura [19] – No correlación entre zonas frías y puntos calientes en los SOM.

Será, por tanto, preciso considerar estos factores para identificar correctamente las variables que interfieren en la conducta suicida. En análisis preciso de cada variable junto con el estudio de los discriminantes, facilitará la selección de los factores de riesgo más importantes.

2.7.2. Discriminantes

Otro criterio relativo a la selección de variables de interés está relacionado con los discriminantes. En el estudio de este proyecto se han desarrollado tres métodos que ayudan a la diferenciación de factores de riesgo.

El discriminante de Fisher es un modelo matemático lineal estudiado en estadística para la toma de decisiones. Si bien, además de este criterio, se han utilizado dos discriminantes adicionales implementados a partir de los parámetros dados por los mapas autoorganizados.

2.7.2.1. Discriminante de Fisher

Este criterio de decisión está implementado al margen de los resultados proporcionados por los mapas autoorganizados. Únicamente se basa en los datos de entrada para calcular los parámetros estadísticos y determinar, para cada variable, un valor discriminante.

Para el desarrollo de estas operaciones se ha implementado una nueva función denominada `som_dFisher.m`. El código recibe como parámetros una matriz del conjunto de variables de estudio con excepción de la variable suicida, que se integra de manera independiente en un vector, así como las etiquetas identificativas de los datos de entrada.

La expresión que implementa el código se corresponde con la representada en la Ecuación [11]. El cálculo de la media y la desviación típica supone, como primer paso, la distinción entre muestras positivas y negativas. Esta clasificación se realiza comprobando, para cada perfil, si la variable suicida es positiva (1) o negativa (0). Cada subgrupo se almacenará en una nueva matriz. En la Figura [20] se representa la matriz de datos y el vector que define el comportamiento suicida. El proceso de división pasa por agrupar aquellas muestras correspondientes a una conducta suicida y definir otro grupo para los perfiles no asociados a intentos de suicidio, tal y como se muestra en la Figura [21].

Variables de Estudio					Variable Suicida
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	...	$x_{1,610}$	1
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	...	$x_{2,610}$	1
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	...	$x_{3,610}$	0
$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	...	$x_{4,610}$	1
		⋮			
$x_{8699,1}$	$x_{8699,2}$	$x_{8699,3}$	...	$x_{8699,610}$	0

Figura [20] – Conjunto de datos de estudio y variable suicida.

$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	...	$x_{1,610}$	1
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	...	$x_{2,610}$	1
$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	...	$x_{4,610}$	1

$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	...	$x_{3,610}$	0
		⋮			
$x_{8699,1}$	$x_{8699,2}$	$x_{8699,3}$	...	$x_{8699,610}$	0

Figura [21] – Subgrupos de muestras positivas (izquierda) y negativas (derecha).

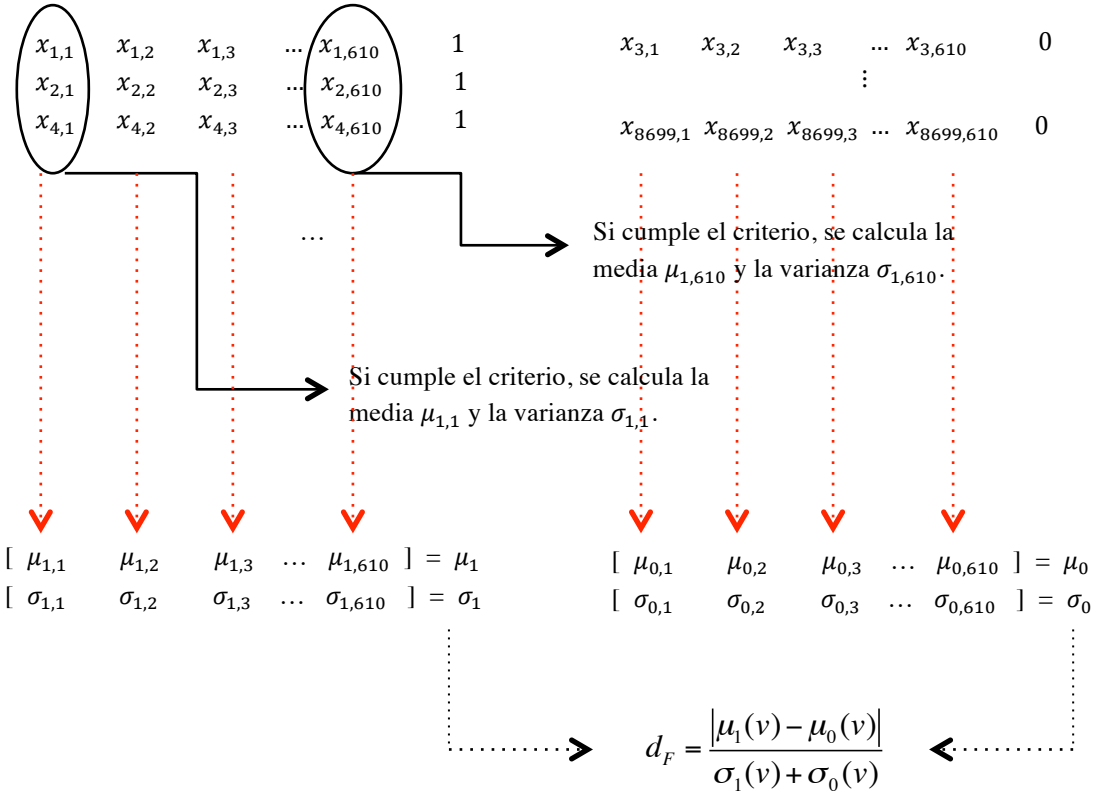


Figura [22] – Proceso de cálculo de medias y varianzas de los subgrupos de muestras.

Sin embargo, esta distinción no resulta tan simple, ya que la presencia de valores perdidos en los datos de estudio supone un problema para los cálculos del programa. La no identificación de valores numéricos no puede ser tratada correctamente por el código, por lo que estos datos no deben tenerse en cuenta en la configuración de los dos subgrupos a partir de los cuales se determinará la media y la varianza.

Para dar fiabilidad a los subgrupos de datos con los que se van a trabajar, será necesario fijar un umbral mínimo para tener en cuenta o no determinadas variables. Es decir, si para una variable de estudio, la mayor parte de muestras se corresponden con valores perdidos, este factor no sería representativo para el estudio. Como mínimo se exige que cada variable cuente con un porcentaje de valores conocidos. De esta forma, se dispondrá de un número de datos suficientes como para determinar la media y la varianza de cada uno de los subgrupos.

El cálculo de los parámetros estadísticos se lleva a cabo recorriendo la matriz de datos e identificando los perfiles suicidas y no suicidas tal y como se indica en la Figura [21]. A partir de los dos subgrupos, se analizará cada una de las variables de manera independiente. Si cumplen con el criterio de porcentaje mínimo de datos definidos, se calculará la media y varianza para cada una de ellas. Este proceso se representa gráficamente en la Figura [22].

2.7.2.2. Discriminante de Fisher aplicado a SOM

Con el objetivo de justificar el uso de mapas autoorganizados, se ha diseñado un nuevo discriminante basado en el criterio de Fisher. Este método de discriminación implementa la función definida por la Ecuación [12]. El primer término del numerador de la expresión hace referencia al vector de centroides en el punto más caliente o *hot spot* del mapa. Cada celda está caracterizada por un vector de las mismas dimensiones que los datos de estudio. Los elementos de estos vectores son, los ya mencionados, centroides. Localizando la celda de mayor riesgo de suicidio, se identificará el vector de centroides asociado a ella. En la Figura [23] se representa el proceso de búsqueda del punto más caliente y de su vector de centroides.

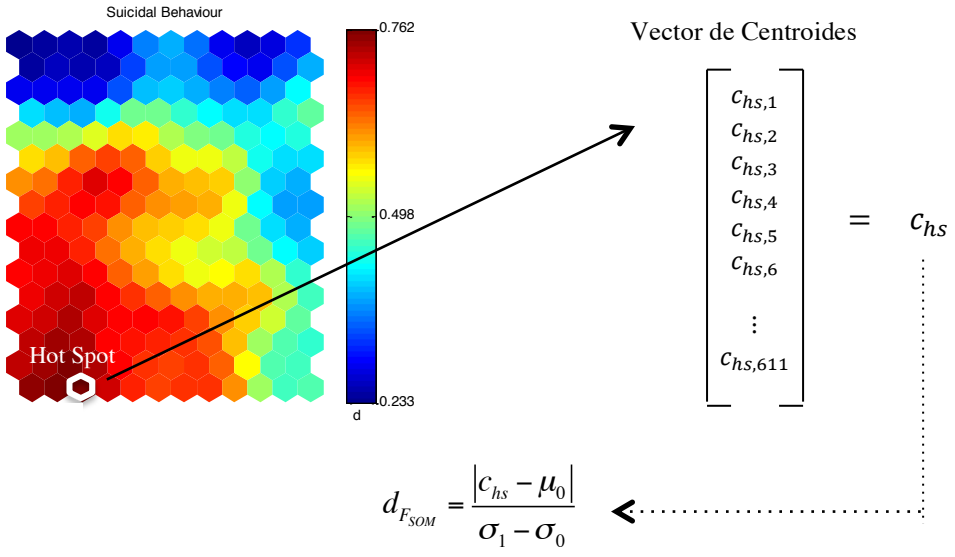


Figura [23] – Identificación del punto más caliente del mapa y de su vector de centroides.

Una vez conocido el centroide más importante, se le restará la media del subgrupo de muestras negativas y el resultado se dividirá por la suma de las desviaciones típicas. Estos parámetros estadísticos se calcularán de nuevo siguiendo el proceso descrito en la Figura [22].

De esta forma, se están involucrando los mapas autoorganizados en la toma de decisiones, validando así este método de clasificación. Se espera, por tanto, que los resultados obtenidos por este discriminante y el anterior sean diferentes, para justificar el uso de esta aplicación en problemas de decisión.

Se ha implementado una nueva función llamada `som_dFisherSOM.m` que recibe como parámetros de entrada las matrices de datos y etiquetas así como los valores de los centroides y el punto más caliente del mapa. Una vez más, vuelve a ser necesario el proceso de subdivisión de las muestras en dos grupos, según si se clasifican como perfiles suicidas o no suicidas. Teniendo en cuenta el problema de los valores no identificados y aplicando las operaciones de la Ecuación [12], a la salida de esta función se obtendrá una lista de todas las variables de estudio ordenadas en orden decreciente, es decir, de mayor a menor relevancia con respecto a la conducta suicida.

2.7.2.3. Discriminante basado en Histogramas

De acuerdo a la expresión de la Ecuación [13], este método de discriminación se basa, principalmente, en los vectores de centroides correspondientes a cada variable así como en los histogramas de muestras positivas y negativas. Es decir, se está buscando una correlación que ayude a destacar las variables de interés.

Para lograr una mayor discriminación, los histogramas de ambos subgrupos deben estar poco solapados, es decir, sus muestras deben distribuirse de manera diferente. Por el contrario, una coincidencia entre ambos histogramas se interpretaría como que no hay una distinción clara entre casos positivos y negativos.

Suponiéndose que los histogramas de subgrupos se distribuyen de manera diferente, es decir, que existen grandes diferencias entre ambos, se analizarían entonces los vectores de centroides de las distintas variables. Si el centroide de mayor valor para una determinada celda coincide con una zona en la que bien el histograma de casos positivos o el histograma de casos negativos alcanza valores elevados, entonces la variable correspondiente a ese centroide se identificará como relevante. Sin embargo, si el centroide destacado para una variable coincide con una zona en que los histogramas se encuentran muy solapados, esa variable será irrelevante para el estudio. La comparativa gráfica entre histogramas y centroides se representa en la Figura [24].

Para la implementación del discriminante basado en histogramas, se ha desarrollado una nueva función denominada `som_dHist.m`, que recibe como parámetros de entrada las matrices de etiquetas y datos de estudio, las dimensiones del mapa, los centroides y los histogramas de casos positivos y negativos.

Para trabajar con los histogramas ha sido precisa una normalización por norma, de modo que todos los valores oscilen entre 0 y 1. A continuación, se calcula la diferencia en valor absoluto entre ambos histogramas, y se multiplica elemento a elemento por el vector de centroides correspondiente a cada una de las variables.

Es decir, para una variable dada, se van recorriendo las celdas del mapa una a una y, de cada una de ellas, se obtiene el centroide relativo a dicha variable, conformándose un nuevo vector de centroides que se multiplicará por la diferencia entre histogramas. Al multiplicar los dos vectores se pretende eliminar aquellos centroides coincidentes con zonas de solapamiento total entre los histogramas. Mediante el cálculo de la diferencia entre histogramas, se estarán anulando las partes en que ambos sean iguales. Los centroides solapados con estas zonas, también se anularán al multiplicarse por cero, de modo que se estarán eliminando muestras no relevantes para el estudio. Valores idénticos de histogramas implican poca distinción entre casos suicidas y no suicidas, lo que dificulta la tarea de identificación y caracterización del comportamiento suicida.

El siguiente paso consiste en realizar el sumatorio del producto para, finalmente, dividir el resultado por la desviación típica correspondiente a la variable en cuestión, tal y como se indica en la Ecuación [13]. Para el cálculo de la varianza, se han omitido los valores perdidos con el objetivo de evitar posibles errores en el código.

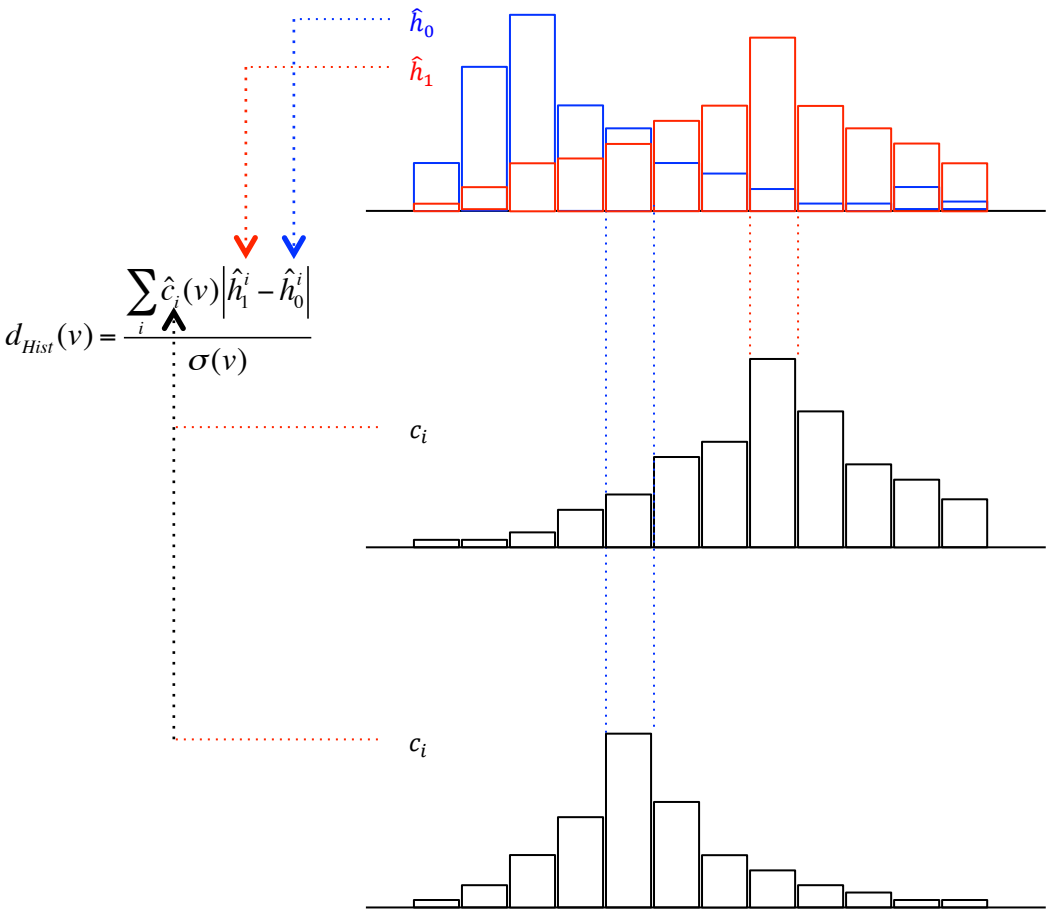


Figura [24] – Solapamiento entre histogramas y centroides.

3. Resultados y Conclusiones

En el apartado 3.1. se realizarán pruebas y analizarán los resultados correspondientes a los diferentes experimentos. En la sección 3.2. se especificarán las conclusiones extraídas en el estudio, mientras que en el apartado 3.3. se explicará cuáles serían las líneas futuras a medio o largo plazo a realizar en el presente proyecto.

3.1. Pruebas y Resultados

En este apartado se analizarán los diferentes experimentos que han sido implementados en el programa. Se estudiarán los parámetros que definen los mapas autoorganizados, como el tipo de inicialización o las dimensiones, además de los criterios discriminantes. Se justificará la toma de decisiones en base a los resultados obtenidos, con la finalidad de alcanzar conclusiones determinantes y que aporten información complementaria a los conocimientos médicos y socioeconómicos ya existentes.

3.1.1. Dimensiones del SOM

Para determinar las dimensiones óptimas de los mapas autoorganizados, se han desarrollado dos criterios basados en las similitudes y diferencias entre los histogramas de casos positivos y negativos.

INICIALIZACIÓN	DIMENSIONES	$\cos\alpha$
Aleatoria	6x2	0,812897
	10x6	0,846822
	16x12	0,851773
	24x20	0,755364
	50x46	**0,675374**
Lineal	6x2	0,820249
	10x6	0,787654
	16x12	0,780404
	24x20	0,772865
	50x46	**0,737192**
Proyección LDA	6x2	0,946696
	10x6	0,820933
	16x12	0,846901
	24x20	0,797062
	50x46	**0,6832010**

Tabla [1] – Coseno del ángulo α formado por los histogramas de casos positivos y negativos.

Con respecto al criterio del coseno, los resultados obtenidos se muestran en la Tabla [1] para los distintos tipos de inicialización y tamaños de mapa. De acuerdo a este método, cuanto menor sea el valor del coseno del ángulo que forman los histogramas de muestras, más diferencias existirán entre ambos y, por tanto, mejor será la distinción entre casos suicidas y no suicidas.

Tras analizar los resultados, se observa que, en general, el valor del coseno disminuye a medida que aumentan las dimensiones del SOM. Estos cambios se interpretan como que, cuanto mayor es el mapa, menos similitudes existen entre los histogramas debido al elevado número de muestras disponibles y, por tanto, mayor es el ángulo entre ambos y, en consecuencia, el valor del coseno estará más próximo a 0. Por el contrario, a menor resolución del mapa, mayor similitud entre muestras e histogramas, que implicará que el ángulo entre ambos sea menor y el valor del coseno tienda a 1 o -1.

INICIALIZACIÓN	DIMENSIONES	DISTANCIA NORMA 1	DISTANCIA NORMA 2
Aleatoria	**6x2**	**0,124779**	**0,050977**
	10x6	0,049514	0,009225
	16x12	0,021985	0,002836
	24x20	0,020943	0,001457
	50x46	0,009650	0,000350
Lineal	**6x2**	**0,133521**	**0,049965**
	10x6	0,056962	0,010861
	16x12	0,031417	0,003452
	24x20	0,020303	0,001404
	50x46	0,008789	0,000315
Proyección LDA	**6x2**	**0,075814**	**0,027209**
	10x6	0,044974	0,009974
	16x12	0,027453	0,002882
	24x20	0,018064	0,001327
	50x46	0,010120	0,000346

Tabla [2] – Distancias de norma 1 y 2 entre los histogramas de casos positivos y negativos.

Con respecto al criterio de la distancia, los resultados obtenidos llevan a la conclusión contraria que en el caso del coseno. Obsérvese la Tabla [2]. Definida la diferencia entre histogramas como la distancia Euclídea entre dos vectores, cuanto mayor sea la distancia, más diferencias existirán entre ambos histogramas. Si bien, ocurre lo contrario que con el criterio del coseno: a menores dimensiones, mayores distancias, y a mayor tamaño, distancias menores.

Por consiguiente, ante la falta de determinación en los resultados de ambos criterios, se define un tamaño intermedio de 16x12 celdas como dimensiones óptimas, ya que en el caso del coseno se prefiere un número de celdas mayor, mientras que para el método de la distancia un tamaño menor devuelve resultados óptimos. Definiéndose un valor intermedio, se alcanza un compromiso entre ambos criterios.

3.1.2. Tipos de Inicialización

Una vez decidido el tamaño óptimo de mapa, se va a proceder al análisis de los distintos tipos de inicialización desarrollados en el proyecto. Se representan cinco gráficas por cada experimento. La primera de ellas hace referencia a la variable suicida (`Suicidal Behaviour`), mientras que las cuatro restantes representan histogramas de muestras.

La imagen `All Cases` grafica el número de sujetos que hay en cada celda del mapa con indiferencia en cuanto a su perfil. Por otro lado, los histogramas `Suicidal Cases` y `Non-Suicidal Cases` representan, respectivamente, los casos de suicidio y no suicidio que hay en cada celda del mapa. Por último, se ha definido un vector de porcentajes para visualizar el tanto por ciento de casos suicidas que hay en cada celda con respecto al total de sujetos de estudio. Esta última gráfica recibe el título `Percentage of Suicidal Cases`.

En las Figuras [25], [26] y [27] se representan los resultados para inicialización aleatoria, lineal y con proyección LDA, respectivamente. Las zonas de color rojo o puntos calientes, representan factores de riesgo para la conducta suicida. Mediante los histogramas de porcentajes puede determinarse la importancia o no de un centroide del mapa dependiendo de la densidad de población que haya en cada celda.

Incialización Aleatoria

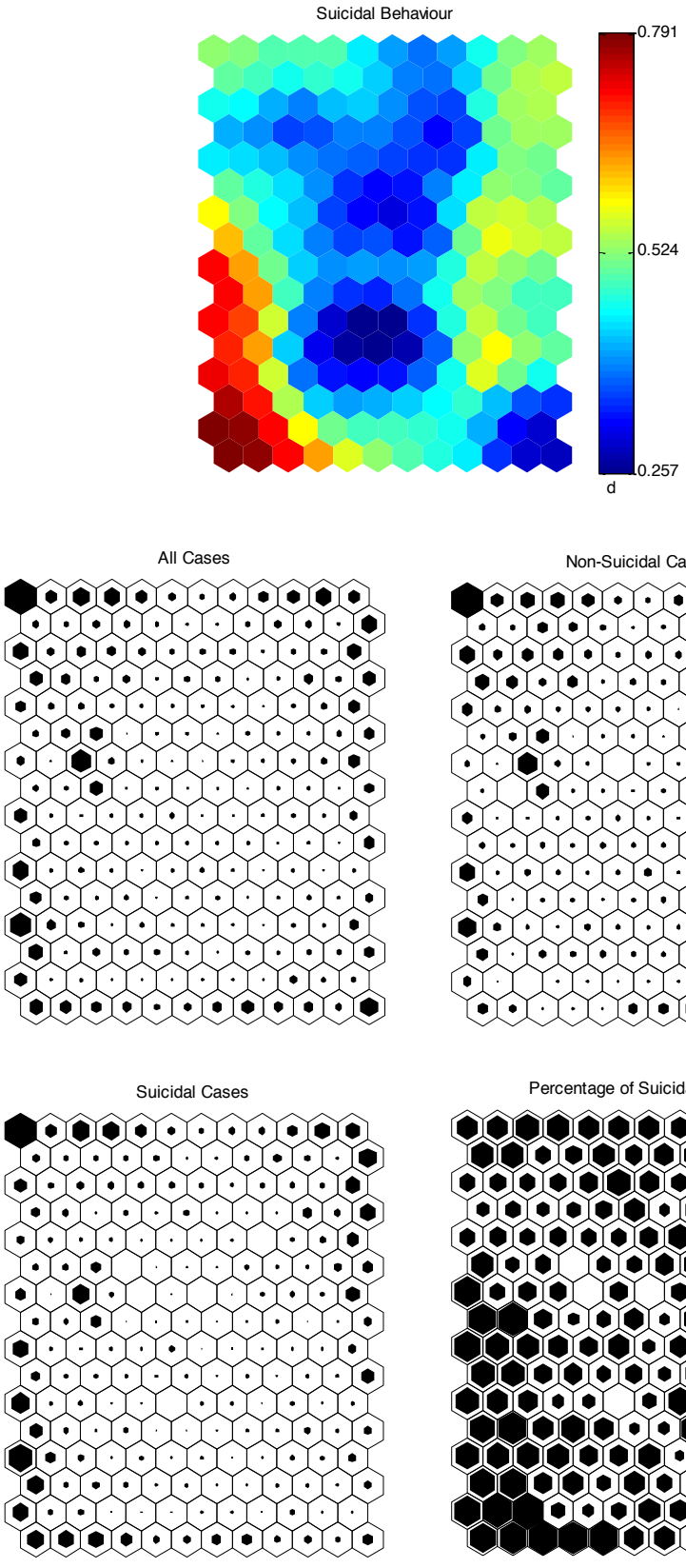
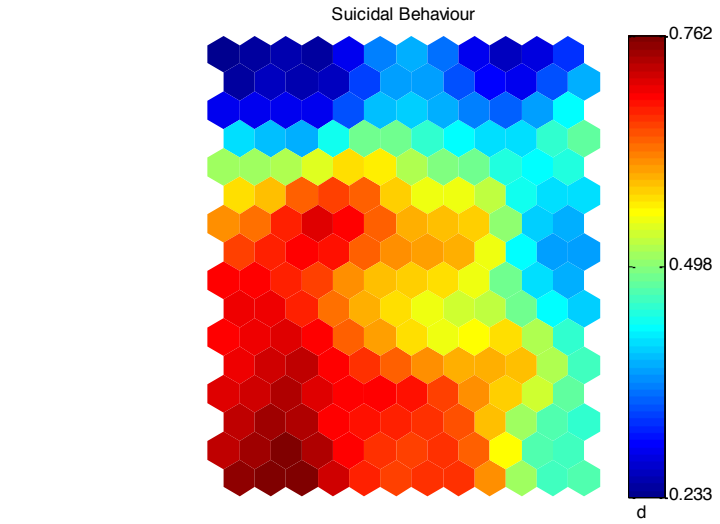


Figura [25] – Gráficas para inicialización aleatoria y dimensiones de16x12 celdas.

Incialización Lineal



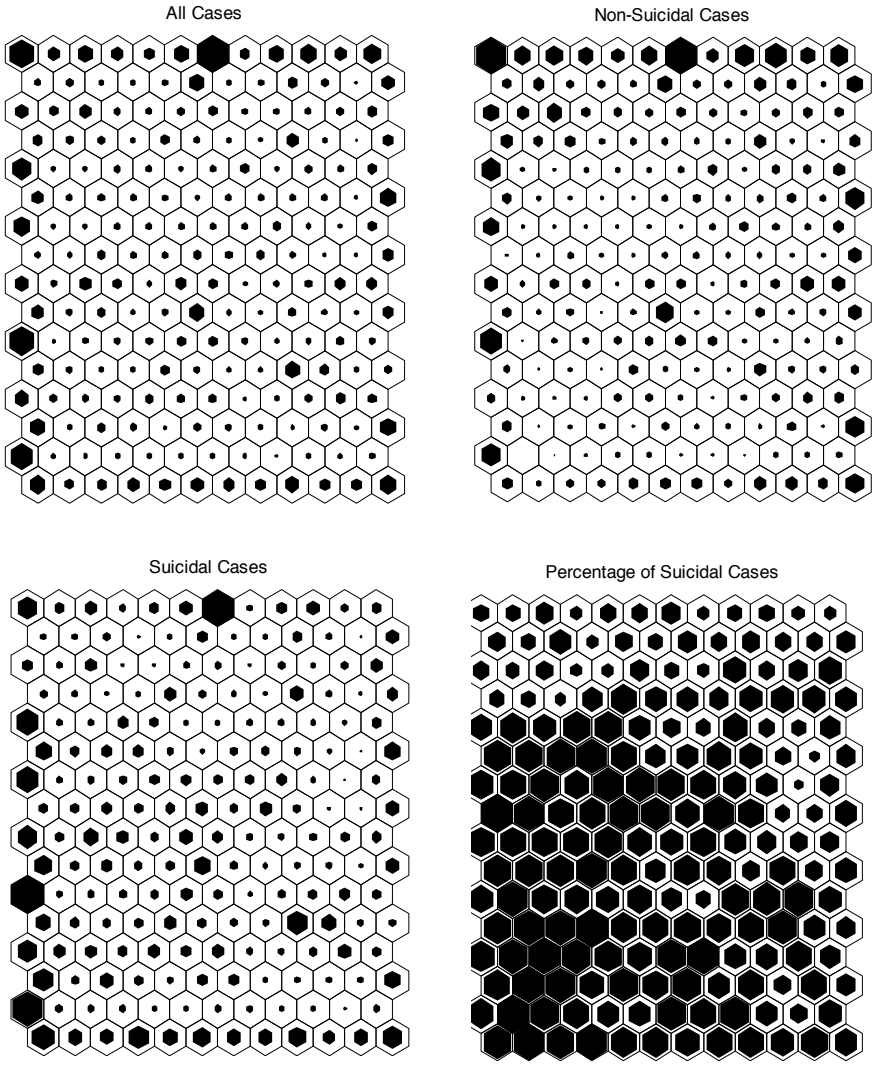


Figura [26] – Gráficas para inicialización lineal y dimensiones de16x12 celdas.

Incialización Proyección LDA

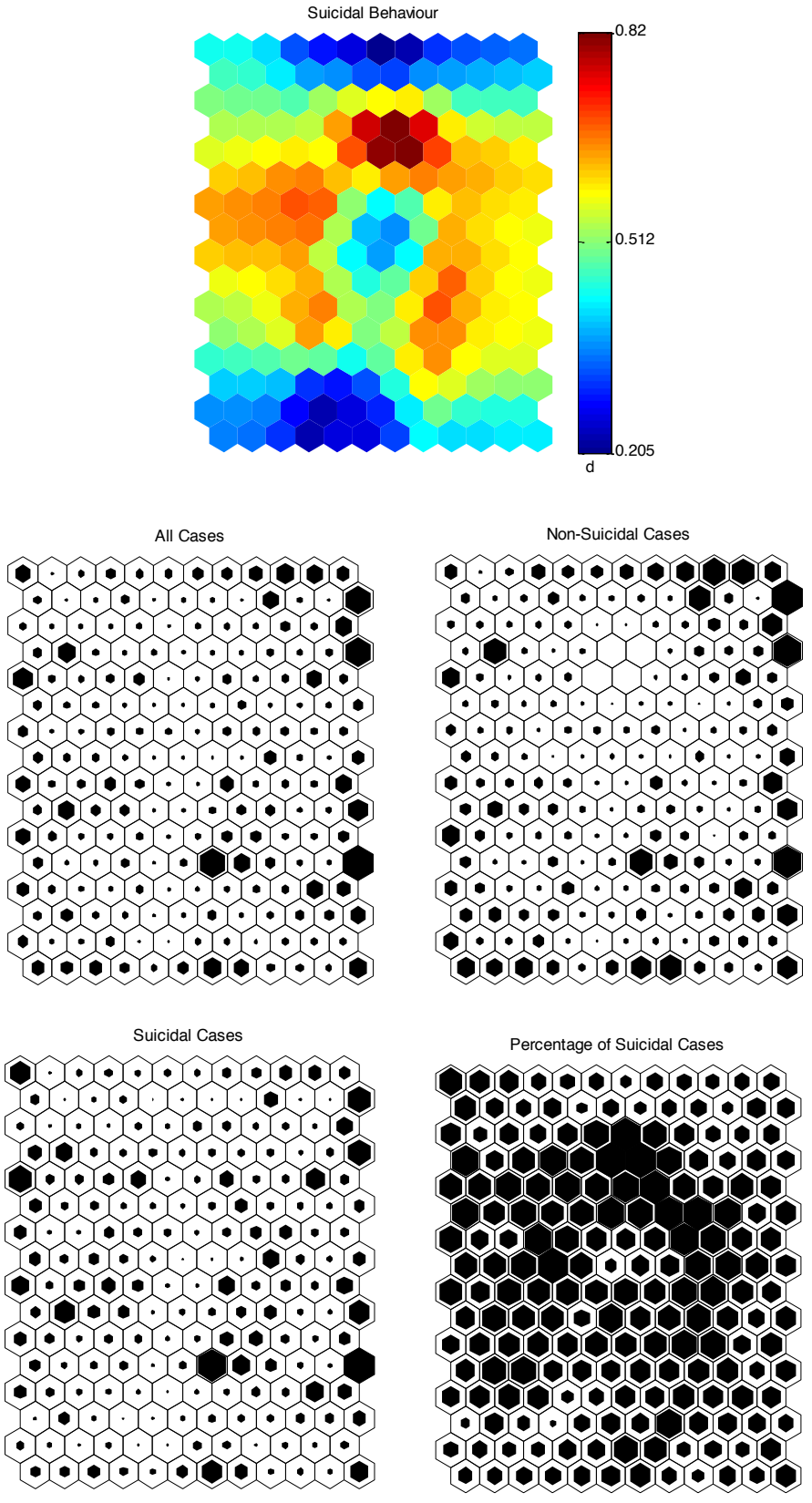


Figura [27] – Gráficas para inicialización con proyección LDA y dimensiones de16x12 celdas.

3.1.3. Porcentajes de Restricción

Antes de determinar las variables de interés de acuerdo a los tres criterios discriminantes para los distintos métodos de inicialización implementados en el programa, deberá tenerse en cuenta otro parámetro. Tal y como se ha mencionado anteriormente, es importante restringir el tanto por ciento de valores perdidos con el que se va a trabajar. Los valores indefinidos restan fiabilidad al estudio, por lo que se necesita limitar su uso para disponer de un mínimo de muestras con información relevante.

Partiendo de un porcentaje de restricción inicial del 15%, se irá incrementando este valor y comparándose los resultados hasta determinar el tanto por ciento óptimo. Es decir, se está asumiendo que, en principio, con un 15% de muestras disponibles, los resultados obtenidos serán fiables. Las pruebas se han realizado aumentando este porcentaje hasta un 40%.

Será preciso aplicar esta restricción sobre el discriminante de Fisher aplicado a SOM, ya que depende tanto de los datos de entrada como de los mapas autoorganizados y, a diferencia del discriminante de Fisher, interviene en la toma de decisiones y en la extracción de resultados. Es decir, el discriminante de Fisher, aunque se calcule también a partir de la media y la desviación típica de las muestras de estudio, se utiliza únicamente para comparar los resultados con el discriminante de Fisher aplicado a SOM y, justificar así, el uso de mapas autoorganizados para la caracterización del comportamiento suicida. Por otro lado, aunque en el discriminante basado en histogramas intervenga la desviación típica, no se tiene en cuenta la media de los datos, luego el discriminante de Fisher aplicado a SOM es el que reúne todos los criterios para analizar consistentemente los distintos tipos de porcentajes.

De este modo, para los porcentajes de restricción seleccionados, se muestran las tablas correspondientes al discriminante de Fisher aplicado a SOM con los distintos métodos de inicialización. En las Tablas [3], [4] y [5] se muestran los resultados para un porcentaje de restricción del 15%. A su vez, las Tablas [6], [7] y [8] reflejan el listado de variables obtenido para una restricción del 20%, mientras que en las Tablas [9], [10] y [11] se incluyen los resultados para un porcentaje restrictivo del 30%. Por último, las Tablas [12], [13] y [14] reflejan las variables de interés para una restricción del 40%.

Porcentaje de Restricción del 15%

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
ctq28_2	0,6525	0,2884	0,3245	0,3748	0,5207
ctq28_14	0,4209	0,1187	0,2501	0,3601	0,4953
gender	0,2387	0,6623	0,4730	0,4918	0,4391
tm	0,9664	0,6155	0,4866	0,3722	0,4086
ctq28_7	0,5621	0,2837	0,3308	0,3651	0,4000
ctq28_19	0,5274	0,2713	0,3256	0,3581	0,3746
ctq28_3	0,3515	0,1379	0,2629	0,3531	0,3468
ctq28_28	0,5042	0,2651	0,3205	0,3736	0,3445
ctq28_25	0,3522	0,1251	0,2748	0,3887	0,3423
his_fam_suicide_behavior	0,3518	0,1152	0,3193	0,4495	0,3078

Tabla [3] – Discriminante $d_{F_{SOM}}$ con inicialización aleatoria y restricción del 15%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
ctq28_14	0,5978	0,1187	0,2501	0,3601	0,7851
ctq28_25	0,6030	0,1251	0,2748	0,3887	0,7204
ctq28_3	0,5508	0,1379	0,2629	0,3531	0,6703
ctq28_24	0,3378	0,0464	0,1727	0,3038	0,6117
ctq28_20	0,3584	0,0582	0,1905	0,3130	0,5964
gender	0,1313	0,6623	0,4730	0,4918	0,5505
ctq28_11	0,4260	0,1189	0,2585	0,3203	0,5305
his_fam_suicide_behavior	0,4914	0,1152	0,3193	0,4495	0,4894
ctq28_23	0,2722	0,0506	0,1782	0,2897	0,4736
ctq28_21	0,2273	0,0325	0,1483	0,2665	0,4694

Tabla [4] – Discriminante $d_{F_{SOM}}$ con inicialización lineal y restricción del 15%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
ctq28_7	0,7783	0,2837	0,3308	0,3651	0,7108
ctq28_2	0,7112	0,2884	0,3245	0,3748	0,6047
gender	0,1172	0,6623	0,4730	0,4918	0,5650
ctq28_11	0,4315	0,1189	0,2585	0,3203	0,5401
dd_hallucinogen	0,1325	0,0087	0,0930	0,1396	0,5325
dd_others	0,1325	0,0138	0,1167	0,1137	0,5153
ctq28_20	0,3156	0,0582	0,1905	0,3130	0,5113
ctq28_25	0,4459	0,1251	0,2748	0,3887	0,4835
niv_edu	0,6832	0,3746	0,3237	0,3536	0,4557
his_fam_suicide_behavior	0,4515	0,1152	0,3193	0,4495	0,4375

Tabla [5] – Discriminante $d_{F_{SOM}}$ con inicialización LDA y restricción del 15%.

Para un porcentaje del 15% se observa la presencia mayoritaria de variables del tipo CTQ, que representan traumas infantiles. Por otro lado, también son importantes las variables género (`gender`) y antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`). En inicialización aleatoria, otra variable destacada es la referente a los trastornos mentales (`tm`), que ocupa la cuarta posición en la tabla. En inicialización LDA intervienen variables relativas a abusos como alucinógenos (`dd_hallucinogen`) y otros (`dd_others`), además del nivel educativo (`niv_edu`).

En el caso de restricción del 20%, las variables género (`gender`), antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`), nivel educarivo (`niv_edu`), año de nacimiento (`yearofbirth_b`), divorcio (`EST_CIV_3`), abuso conjunto de alcohol y drogas (`dd_al_drug`) y trastornos mentales como depresión (`dd_depre`) y ansiedad (`dd_anxiety`) están presentes en los tres tipos de inicialización. Otros factores importantes coincidentes en dos de los experimentos son el abuso de alcohol (`dd_oh`) y el abuso de cannabis (`dd_cannabis`).

Porcentaje de Restricción del 20%

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,2387	0,6623	0,4730	0,4918	0,4391
his_fam_suicide_behavior	0,3518	0,1152	0,3193	0,4495	0,3078
dd_anxiety	0,0848	0,0285	0,1169	0,1245	0,2334
dd_depre	0,0757	0,0325	0,1069	0,1101	0,1995
dd_psychotic	0,1565	0,0682	0,2521	0,3791	0,1400
niv_edu	0,4658	0,3746	0,3237	0,3536	0,1346
yearofbirth_b	0,7068	0,6666	0,1496	0,1504	0,1339
EST_CIV_3	0,1938	0,1068	0,3089	0,3541	0,1312
dd_al_drug	0,2110	0,3325	0,4712	0,4791	0,1279
dd_substance	0,1204	0,2189	0,4136	0,4029	0,1206

Tabla [6] - Discriminante $d_{F_{SOM}}$ con inicialización aleatoria y restricción del 20%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1313	0,6623	0,4730	0,4918	0,5505
his_fam_suicide_behavior	0,4914	0,1152	0,3193	0,4495	0,4894
dd_oh	0,4349	0,1186	0,3234	0,4412	0,4136
dd_al_drug	0,6380	0,3325	0,4712	0,4791	0,3214
niv_edu	0,5450	0,3746	0,3237	0,3536	0,2517
dd_anxiety	0,0881	0,0285	0,1169	0,1245	0,2472
yearofbirth_b	0,7303	0,6666	0,1496	0,1504	0,2123
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1887
EST_CIV_3	0,2151	0,1068	0,3089	0,3541	0,1633
dd_cannabis	0,1495	0,0571	0,2321	0,3354	0,1629

Tabla [7] - Discriminante $d_{F_{SOM}}$ con inicialización lineal y restricción del 20%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
dd_oh	0,5395	0,1186	0,3234	0,4412	0,5505
EST_CIV_3	0,4313	0,1068	0,3089	0,3541	0,4894
niv_edu	0,6547	0,3746	0,3237	0,3536	0,4136
his_fam_suicide_behavior	0,3623	0,1152	0,3193	0,4495	0,3214
dd_al_drug	0,5717	0,3325	0,4712	0,4791	0,2517
dd_anxiety	0,0881	0,0285	0,1169	0,1245	0,2469
yearofbirth_b	0,7303	0,6666	0,1496	0,1504	0,2123
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1885
gender	0,8199	0,6623	0,4730	0,4918	0,1633
dd_cannabis	0,1495	0,0571	0,2321	0,3354	0,1628

Tabla [8] - Discriminante $d_{F_{SOM}}$ con inicialización LDA y restricción del 20%.

Porcentaje de Restricción del 30%

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,2387	0,6623	0,4730	0,4918	0,4391
his_fam_suicide_behavior	0,3518	0,1152	0,3193	0,4495	0,3078
dd_anxiety	0,0848	0,0285	0,1169	0,1245	0,2334
dd_depre	0,0757	0,0325	0,1069	0,1101	0,1995
niv_edu	0,4658	0,3746	0,3237	0,3536	0,1346
EST_CIV_3	0,1938	0,1068	0,3089	0,3541	0,1312
dd_al_drug	0,2110	0,3325	0,4712	0,4791	0,1279
dd_substance	0,1204	0,2189	0,4136	0,4029	0,1206
dd_bipolar	0,0353	0,0155	0,0862	0,0983	0,1073
EST_CIV_1	0,3046	0,3658	0,4817	0,4645	0,0647

Tabla [9] - Discriminante $d_{F_{SOM}}$ con inicialización aleatoria y restricción del 30%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1313	0,6623	0,4730	0,4918	0,5505
his_fam_suicide_behavior	0,4914	0,1152	0,3193	0,4495	0,4894
dd_oh	0,4349	0,1186	0,3234	0,4412	0,4136
dd_al_drug	0,6380	0,3325	0,4712	0,4791	0,3214
niv_edu	0,5450	0,3746	0,3237	0,3536	0,2517
dd_anxiety	0,0881	0,0285	0,1169	0,1245	0,2472
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1887
EST_CIV_3	0,2151	0,1068	0,3089	0,3541	0,1633
dd_bipolar	0,0412	0,0155	0,0862	0,0983	0,1397
age	0,2867	0,3284	0,1480	0,1573	0,1368

Tabla [10] - Discriminante $d_{F_{SOM}}$ con inicialización lineal y restricción del 30%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
dd_oh	0,5506	0,1186	0,3234	0,4412	0,5650
niv_edu	0,6832	0,3746	0,3237	0,3536	0,4557
his_fam_suicide_behavior	0,4515	0,1152	0,3193	0,4495	0,4375
EST_CIV_2	0,1367	0,3176	0,4656	0,4317	0,2016
EST_CIV_3	0,2271	0,1068	0,3089	0,3541	0,1814
EST_CIV_1	0,1989	0,3658	0,4817	0,4645	0,1764
dd_depre	0,0698	0,0325	0,1069	0,1101	0,1723
dd_al_drug	0,1953	0,3325	0,4712	0,4791	0,1443
age	0,2890	0,3284	0,1480	0,1573	0,1291
gender	0,7686	0,6623	0,4730	0,4918	0,1102

Tabla [11] - Discriminante $d_{F_{SOM}}$ con inicialización LDA y restricción del 30%.

Aplicando una restricción del 30% destacan como factores de interés el género (`gender`), los antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`), el nivel educativo (`niv_edu`), el divorcio (`EST_CIV_3`), la depresión (`dd_depre`) y el abuso combinado de alcohol y drogas (`dd_al_drug`).

Para un porcentaje del 40%, sólo existen siete variables que cumplan esa restricción: el género (`gender`), la edad (`age`), la depresión (`dd_depre`) y los cuatro estados civiles definidos, soltero (`EST_CIV_1`), casado (`EST_CIV_2`), divorciado (`EST_CIV_3`) o viudo (`EST_CIV_4`). En función del tipo de inicialización, cambia el orden de relevancia de estos factores.

Porcentaje de Restricción del 40%

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,2387	0,6623	0,4730	0,4918	0,4391
dd_depre	0,0757	0,0325	0,1069	0,1101	0,1995
EST_CIV_3	0,1938	0,1068	0,3089	0,3541	0,1312
EST_CIV_1	0,3046	0,3658	0,4817	0,4645	0,0647
EST_CIV_2	0,2650	0,3176	0,4656	0,4317	0,0587
age	0,3140	0,3284	0,1480	0,1573	0,0472
EST_CIV_4	0,0139	0,0215	0,1451	0,1354	0,0272

Tabla [12] - Discriminante $d_{F_{SOM}}$ con inicialización aleatoria y restricción del 40%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1313	0,6623	0,4730	0,4918	0,5504
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1887
EST_CIV_3	0,2151	0,1068	0,3089	0,3541	0,1633
age	0,2867	0,3284	0,1480	0,1573	0,1368
EST_CIV_1	0,2783	0,3658	0,4817	0,4645	0,0925
EST_CIV_2	0,2684	0,3176	0,4656	0,4317	0,0548
EST_CIV_4	0,0109	0,0215	0,1451	0,1354	0,0379

Tabla [13] - Discriminante $d_{F_{SOM}}$ con inicialización lineal y restricción del 40%.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1172	0,6623	0,4730	0,4918	0,5650
EST_CIV_2	0,1367	0,3176	0,4656	0,4317	0,2016
EST_CIV_3	0,2271	0,1068	0,3089	0,3541	0,1814
EST_CIV_1	0,1989	0,3658	0,4817	0,4645	0,1764
dd_depre	0,0698	0,0325	0,1069	0,1101	0,1723
age	0,2890	0,3284	0,1480	0,1573	0,1291
EST_CIV_4	0,0008	0,0215	0,1451	0,1354	0,0740

Tabla [14] - Discriminante $d_{F_{SOM}}$ con inicialización LDA y restricción del 40%.

Debido a que en este último caso la restricción es muy fuerte y a que en el primero de los experimentos, quizás, la restricción sea muy débil obteniéndose mayoritariamente variables del tipo CTQ, se descartan ambos porcentajes como valores restrictivos óptimos.

Analizando las otras dos restricciones, ambas serían igualmente válidas, teniéndose en los dos casos conjuntos de variables coherentes para el estudio. Sin embargo, optando por un criterio que no sea demasiado restrictivo para no eliminar información útil y evitar restar fiabilidad a los resultados, se decide aplicar un porcentaje del 20% en los sucesivos experimentos del proyecto.

3.1.4. Tablas de Discriminantes

Para cada criterio discriminante, se va a realizar un experimento con la finalidad de obtener un listado con las diez variables más importantes que resulten de interés para el comportamiento suicida. En base a cada discriminante, se analizará cada uno de los tres métodos de inicialización implementados en el código, aplicando un porcentaje de restricción del 20% en los casos de Fisher y Fisher aplicado a SOM.

3.1.4.1. Discriminante de Fisher

En el caso del discriminante de Fisher, puesto que éste sólo depende de los estadísticos de los datos de entrada, y no de los parámetros de los mapas autoorganizados, es indiferente el método de inicialización que se aplique. Por tanto, para los tres tipos de inicialización, los resultados serán los mismos en cualquier caso.

Bajo este criterio no es necesario extraer conclusiones relevantes para el estudio, puesto que este discriminante es puramente comparativo, es decir, es una referencia con la que comparar los resultados del discriminante de Fisher aplicado a SOM y, de esta manera, justificar la aplicación de los mapas autoorganizados para la caracterización del comportamiento suicida.

Variable	μ_0	μ_1	σ_0	σ_1	d_{Fisher}
gender	0,6623	0,4095	0,4730	0,4918	0,2621
niv_edu	0,3746	0,5431	0,3237	0,3536	0,2487
his_fam_suicide_behavior	0,1152	0,2808	0,3193	0,4495	0,2154
dd_oh	0,1186	0,2647	0,3234	0,4412	0,1910
dd_psychotic	0,0682	0,1739	0,2521	0,3791	0,1675
dd_anxiety	0,0285	0,0686	0,1169	0,1245	0,1663
dd_depre	0,0325	0,0665	0,1069	0,1101	0,1570
dd_cannabis	0,0571	0,1291	0,2321	0,3354	0,1268
yearofbirth_b	0,6666	0,6900	0,1496	0,1504	0,0781
EST_CIV_2	0,3176	0,2476	0,4656	0,4317	0,0780

Tabla [15] – Discriminante de Fisher con cualquier tipo de inicialización.

Los resultados obtenidos corresponden, en su mayoría, a trastornos mentales como la psicosis (dd_psychotic), la ansiedad (dd_anxiety) y la depresión (dd_depre) y a abusos como alcohol (dd_oh) o cannabis (dd_cannabis). También son importantes variables como el género (gender), el año de nacimiento (yearofbirth_b), el nivel educativo (niv_edu), el estado civil casado (EST_CIV_2) o los antecedentes familiares de conducta suicida (his_fam_suicide_behavior).

3.1.4.2. Discriminante de Fisher aplicado a SOM

En las siguientes tablas se muestran los conjuntos de variables según el criterio del discriminante de Fisher aplicado a SOM para los distintos tipos de inicialización implementados en el programa. Los resultados esperados deberían diferir de los obtenidos por el discriminante de Fisher, reflejados en la Tabla [15].

Las Tablas [16], [17] y [18] incluyen los resultados del discriminante de Fisher aplicado a SOM para inicialización aleatoria, lineal y con proyección LDA, respectivamente.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,2387	0,6623	0,4730	0,4918	0,4391
his_fam_suicide_behavior	0,3518	0,1152	0,3193	0,4495	0,3078
dd_anxiety	0,0848	0,0285	0,1169	0,1245	0,2334
dd_depre	0,0757	0,0325	0,1069	0,1101	0,1995
dd_psychotic	0,1565	0,0682	0,2521	0,3791	0,1400
niv_edu	0,4658	0,3746	0,3237	0,3536	0,1346
yearofbirth_b	0,7068	0,6666	0,1496	0,1504	0,1339
EST_CIV_3	0,1938	0,1068	0,3089	0,3541	0,1312
dd_al_drug	0,2110	0,3325	0,4712	0,4791	0,1279
dd_substance	0,1204	0,2189	0,4136	0,4029	0,1206

Tabla [16] - Discriminante de Fisher aplicado a SOM con inicialización aleatoria.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1313	0,6623	0,4730	0,4918	0,5505
his_fam_suicide_behavior	0,4914	0,1152	0,3193	0,4495	0,4894
dd_oh	0,4349	0,1186	0,3234	0,4412	0,4136
dd_al_drug	0,6380	0,3325	0,4712	0,4791	0,3214
niv_edu	0,5450	0,3746	0,3237	0,3536	0,2517
dd_anxiety	0,0881	0,0285	0,1169	0,1245	0,2472
yearofbirth_b	0,7303	0,6666	0,1496	0,1504	0,2123
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1887
EST_CIV_3	0,2151	0,1068	0,3089	0,3541	0,1633
dd_cannabis	0,1495	0,0571	0,2321	0,3354	0,1629

Tabla [17] - Discriminante de Fisher aplicado a SOM con inicialización lineal.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
dd_oh	0,5395	0,1186	0,3234	0,4412	0,5505
EST_CIV_3	0,4313	0,1068	0,3089	0,3541	0,4894
niv_edu	0,6547	0,3746	0,3237	0,3536	0,4136
his_fam_suicide_behavior	0,3623	0,1152	0,3193	0,4495	0,3214
dd_al_drug	0,5717	0,3325	0,4712	0,4791	0,2517
dd_anxiety	0,0881	0,0285	0,1169	0,1245	0,2469
yearofbirth_b	0,7303	0,6666	0,1496	0,1504	0,2123
dd_depre	0,0734	0,0325	0,1069	0,1101	0,1885
gender	0,8199	0,6623	0,4730	0,4918	0,1633
dd_cannabis	0,1495	0,0571	0,2321	0,3354	0,1628

Tabla [18] – Discriminante de Fisher aplicado a SOM con inicialización LDA.

Tal y como se ha indicado en el apartado anterior, para el discriminante de Fisher aplicado a SOM con una restricción del 20%, las variables más importantes debido a su presencia en las tres inicializaciones son el género (`gender`), los antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`), el nivel educarivo (`niv_edu`), el año de nacimiento (`yearofbirth_b`), el divorcio (`EST_CIV_3`), el abuso conjunto de alcohol y drogas (`dd_al_drug`) y los trastornos mentales como depresión (`dd_depre`) y ansiedad (`dd_anxiety`). Otros factores coincidentes en dos de los experimentos realizados son el abuso de alcohol (`dd_oh`) y el abuso de cannabis (`dd_cannabis`).

3.1.4.3. Discriminante basado en Histogramas

Con respecto a este discriminante, la única restricción aplicada es de un 15% para controlar el porcentaje de valores perdidos implicados en el cálculo de la desviación típica, tal y como se indica en la Ecuación [13].

En cuanto a los conjuntos de variables indicados en las Tablas [19], [20] y [21], puede observarse, en todos los casos, la presencia de variables como el año de nacimiento (`yearofbirth_b`), el estado de ansiedad de acuerdo a los criterios de Molise y Montpellier (`STAI_anxiety`), la hostilidad verbal (`BDHI_Hostivb`), la impulsividad medida según la planificación (`BIS10_impulnp`), la irritabilidad (`BDHI_Irritab`), el control de enfados (`STAXI_Angecontrol`) y la culpabilidad (`BDHI_Culpab`).

Es destacable la importancia adquirida por la variable año de nacimiento que, a priori, no guarda ninguna relación con el comportamiento suicida. En comparación a los valores discriminantes obtenidos para Fisher y Fisher aplicado a SOM, es notable el incremento de magnitud en este caso, a pesar de la normalización aplicada sobre los datos.

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	18,7933
STAI_anxiety	0,1345	14,3216
BDHI_Hostivb	0,1885	13,3905
BIS10_impulnp	0,1581	12,5693
BDHI_Irritab	0,2154	12,4023
STAXI_Angecontrol	0,1986	11,8248
BIS10_Impulc	0,1788	11,6843
BDHI_Culpab	0,2452	11,0456
BDHI_Hostind	0,2202	10,8793
BDHI18	0,3441	10,6506

Tabla [19] – Discriminante basado en histogramas para inicialización aleatoria.

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	26,6727
STAI_anxiety	0,1345	21,6075
BDHI_Hostivb	0,1885	18,1740
STAXI_Angecontrol	0,1986	17,8963
BDHI_Irritab	0,2154	16,9554
BIS10_impulnp	0,1581	16,5471
MO_BIS27	0,2717	16,3763
BDHI_Culpab	0,2452	16,3072
MO_BIS30	0,2681	15,9267
mo_bdhi31	0,3355	15,7749

Tabla [20] – Discriminante basado en histogramas para inicialización lineal.

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	24,0305
STAI_anxiety	0,1345	18,3097
BDHI_Irritab	0,2154	16,6147
BDHI_Hostivb	0,1885	16,4282
BDHI_Culpab	0,2452	15,2798
BIS10_impulnp	0,1581	14,9645
STAXI_Angecontrol	0,1986	14,9091
BDHI_Hostind	0,2202	14,2400
BIS10_Impulc	0,1788	13,9499
mo_bdhi31	0,3355	13,7218

Tabla [21] – Discriminante basado en histogramas para inicialización LDA.

A continuación, se analizarán distintas serie de cuatro gráficas por cada método de inicialización. Existe un conjunto de tres diagramas de barras así como dos histogramas hexagonales para cada caso. El diagrama de barras define, en primer lugar, la distancia entre histogramas de casos positivos y negativos. La segunda distribución hace referencia al vector de centroides correspondiente a una de las cuatro variables más influyentes. Por último, también es importante visualizar el producto entre las dos gráficas anteriores para comprobar el solapamiento de las muestras.

Además, los histogramas hexagonales representan la diferencia entre histogramas positivos y negativos y su producto con los centroides de cada variable. Este último histograma ha sido ecualizado dividiendo cada centroide por el de mayor valor para visualizar de forma más clara la ocupación de cada celda. De esta forma, una de las celdas tendrá valor 1 y será totalmente oscura, quedando definido el centroide más importante para cada variable y haciendo más fácil su identificación.

Inicialización Aleatoria

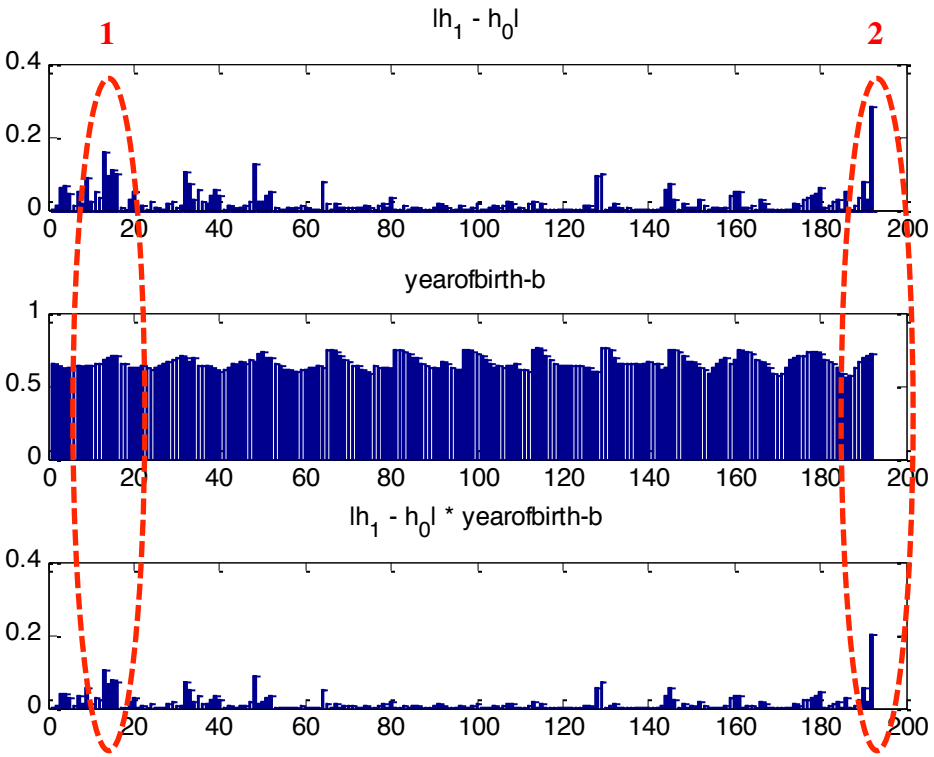


Figura [28] – Diferencia entre histogramas, centroides de la variable `yearofbirth-b` y producto entre ambas gráficas con inicialización aleatoria.

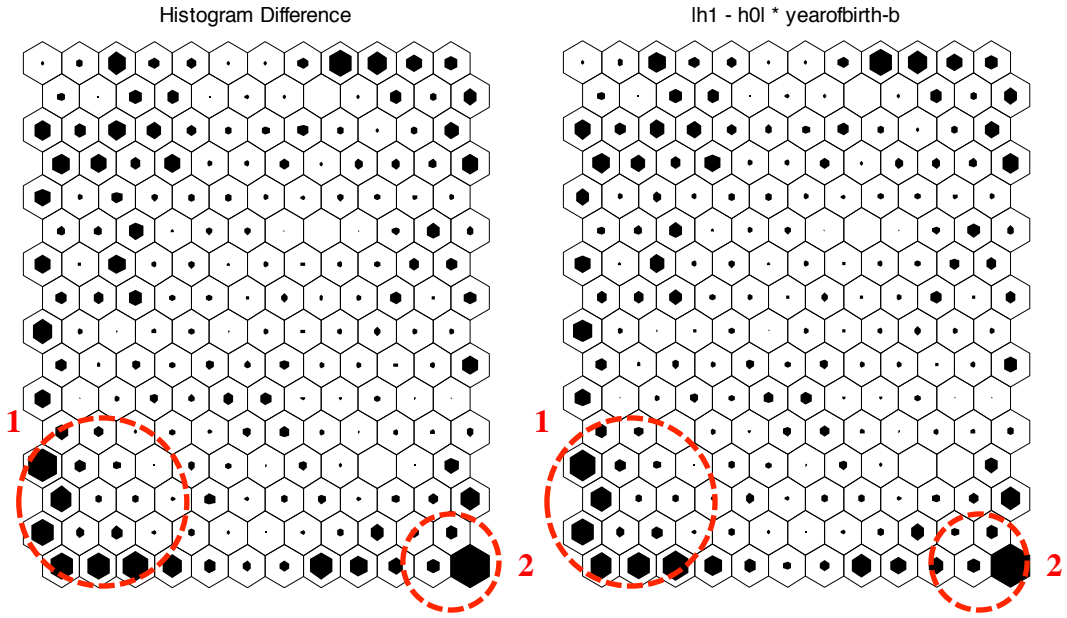


Figura [29] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `yearofbirth-b` con inicialización aleatoria.

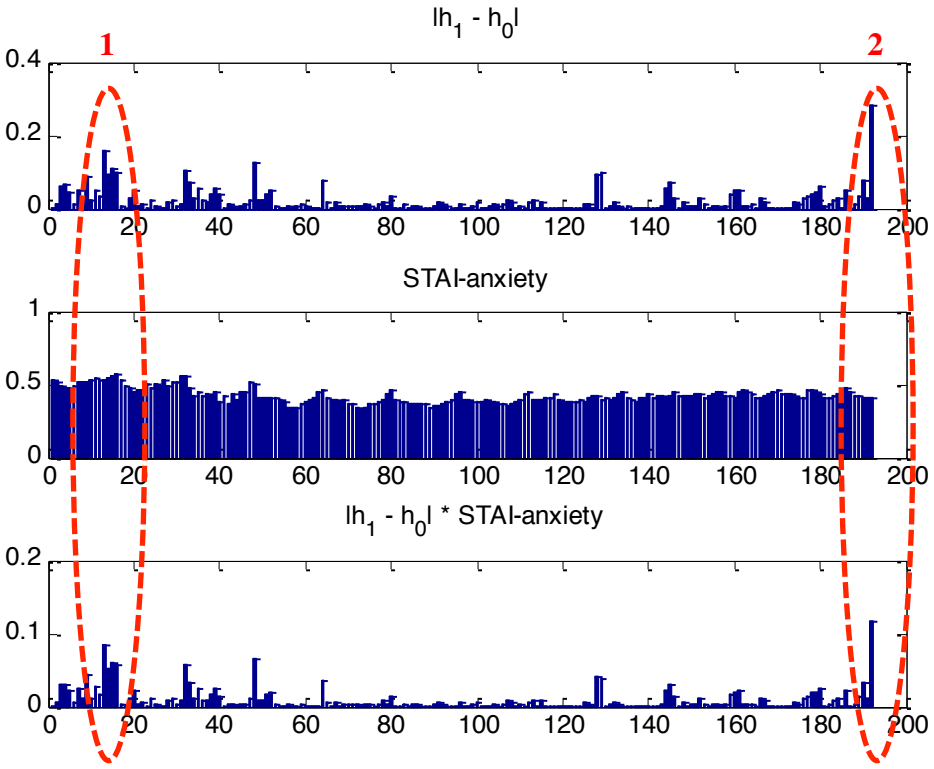


Figura [30] – Diferencia entre histogramas, centroides de la variable `STAI-anxiety` y producto entre ambas gráficas con inicialización aleatoria.

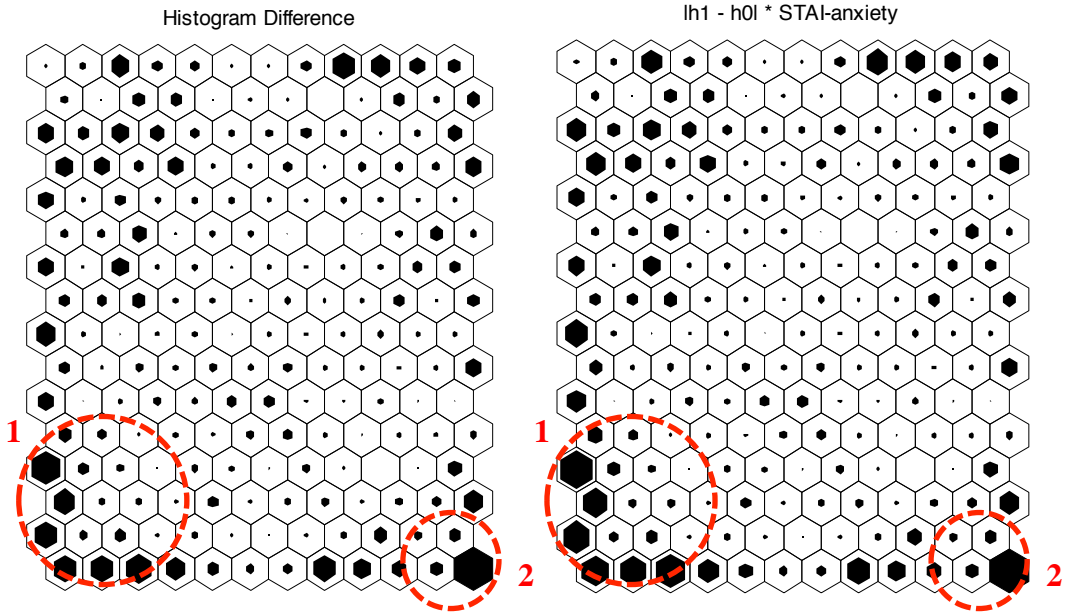


Figura [31] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `STAI-anxiety` con inicialización aleatoria.

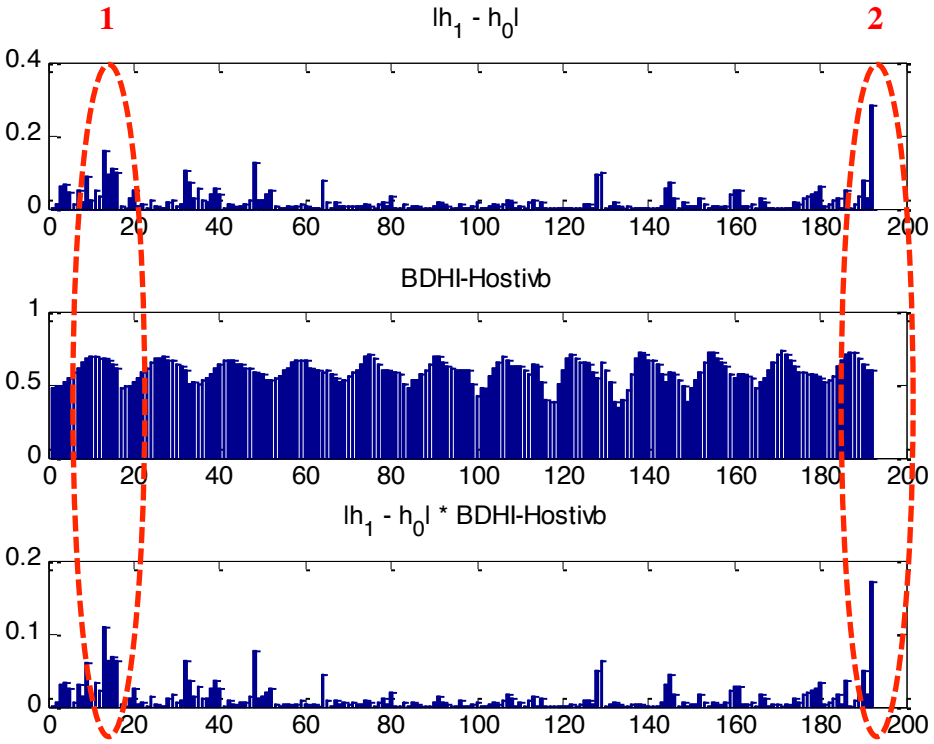


Figura [32] – Diferencia entre histogramas, centroides de la variable `BDHI-Hostivb` y producto entre ambas gráficas con inicialización aleatoria.

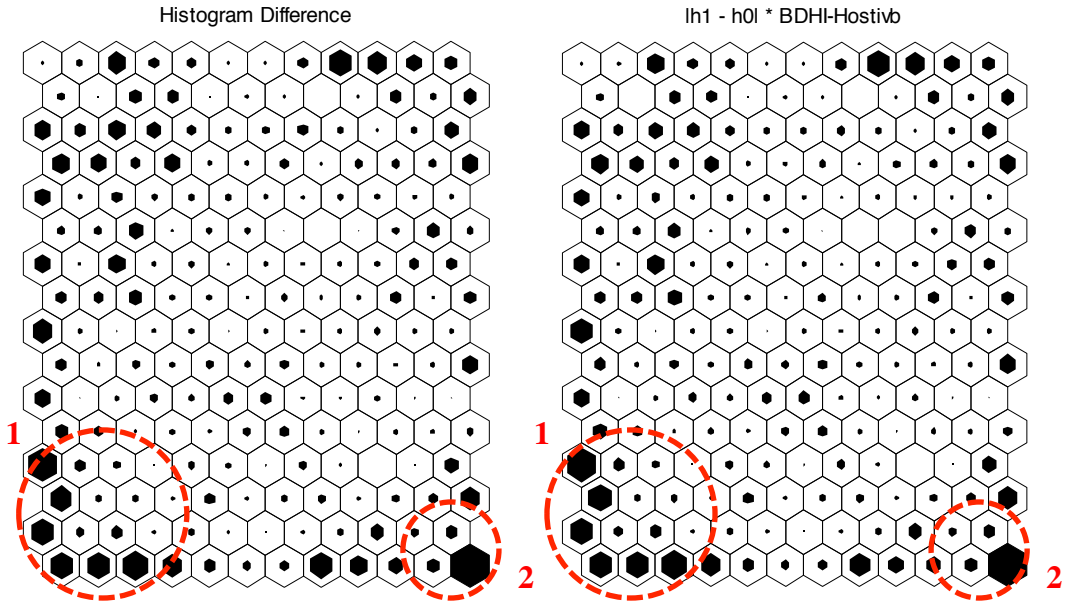


Figura [33] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `BDHI-Hostivb` con inicialización aleatoria.

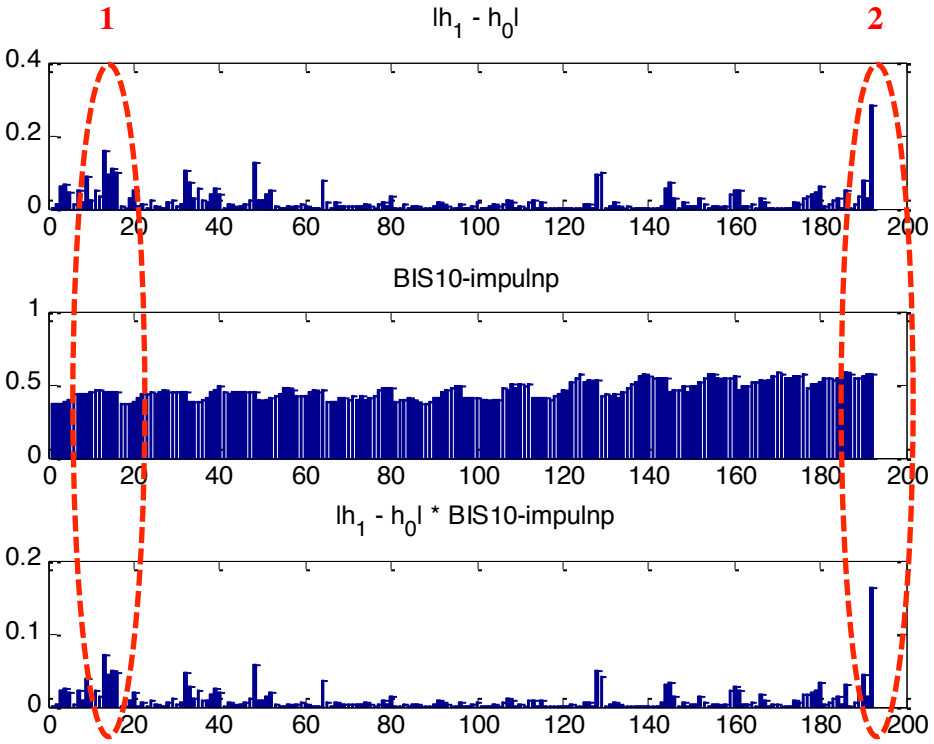


Figura [34] – Diferencia entre histogramas, centroides de la variable `BIS10-impulnp` y producto entre ambas gráficas con inicialización aleatoria.

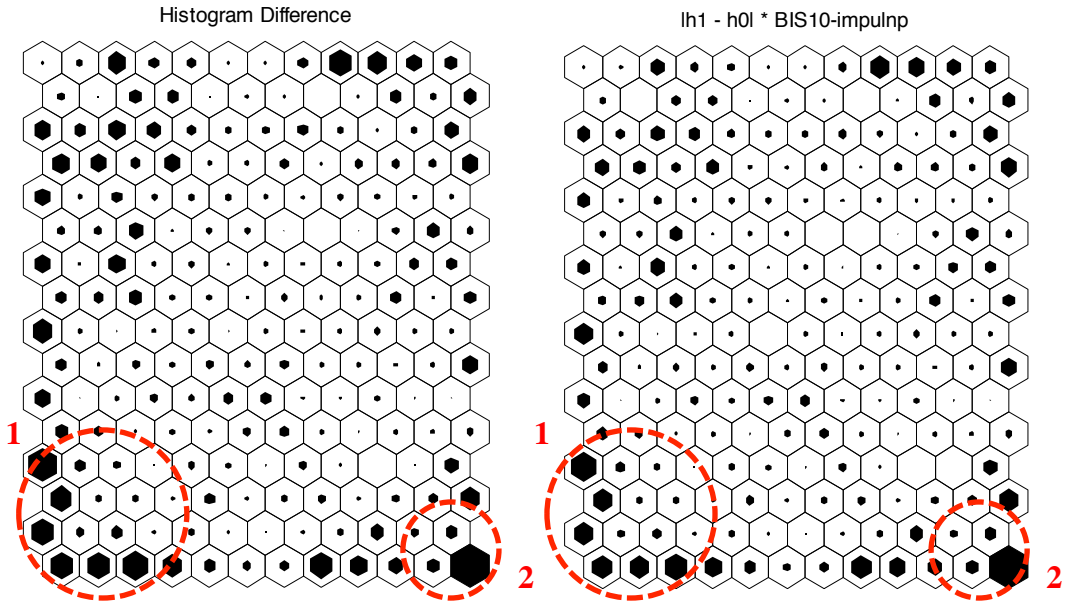


Figura [35] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `BIS10-impulnp` con inicialización aleatoria.

Inicialización Lineal

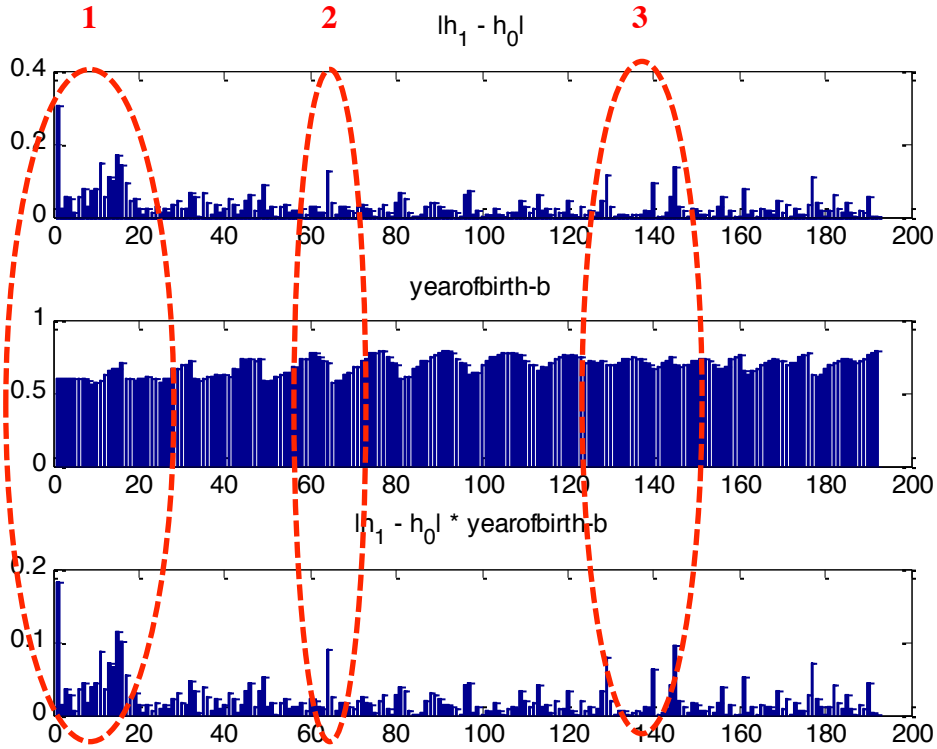


Figura [36] – Diferencia entre histogramas, centroides de la variable `yearofbirth-b` y producto entre ambas gráficas con inicialización lineal.

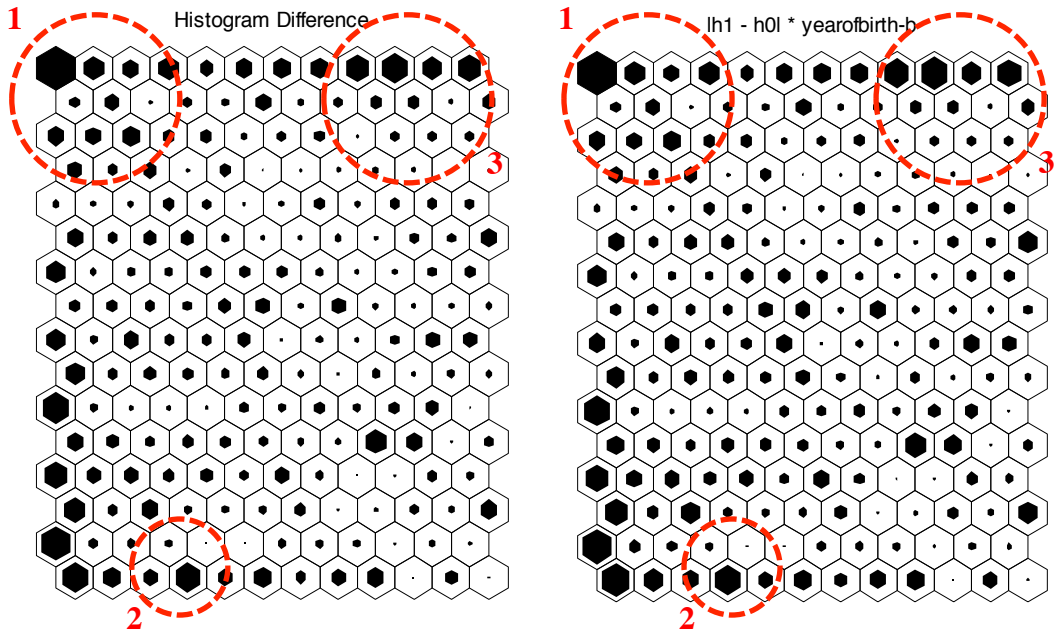


Figura [37] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `yearofbirth-b` con inicialización lineal.

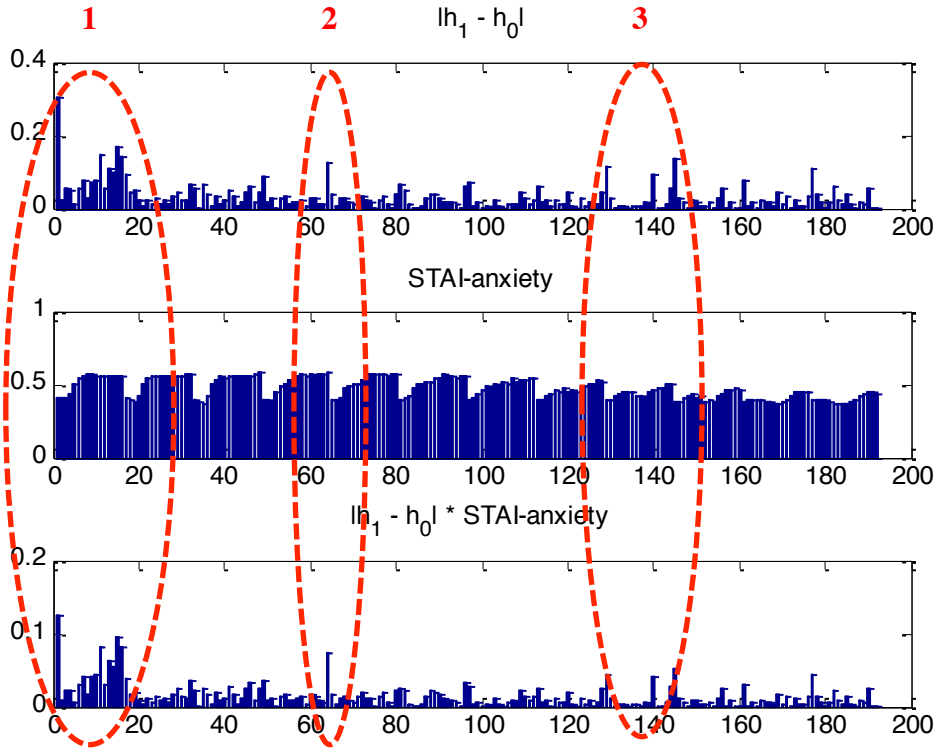


Figura [38] – Diferencia entre histogramas, centroides de la variable `STAI-anxiety` y producto entre ambas gráficas con inicialización lineal.

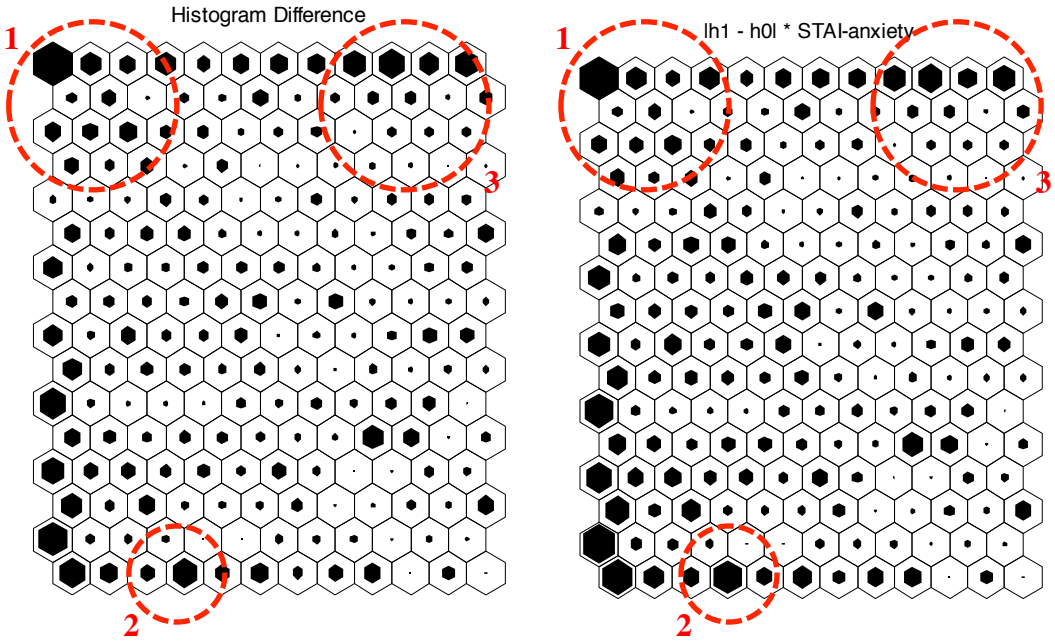


Figura [39] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `STAI-anxiety` con inicialización lineal.

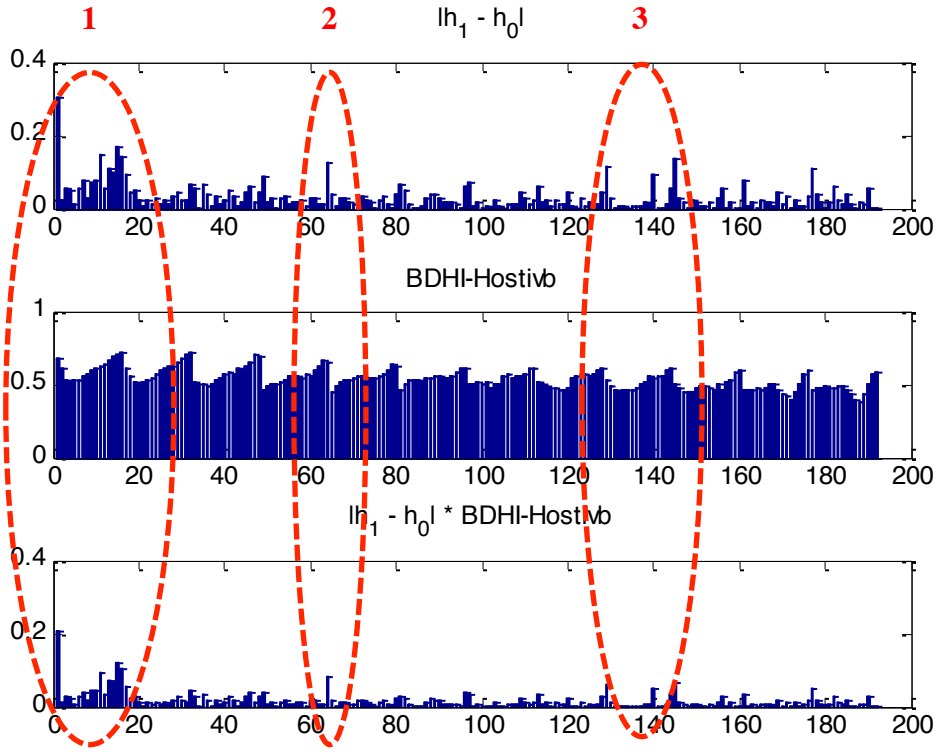


Figura [40] – Diferencia entre histogramas, centroides de la variable `BDHI-Hostivb` y producto entre ambas gráficas con inicialización lineal.

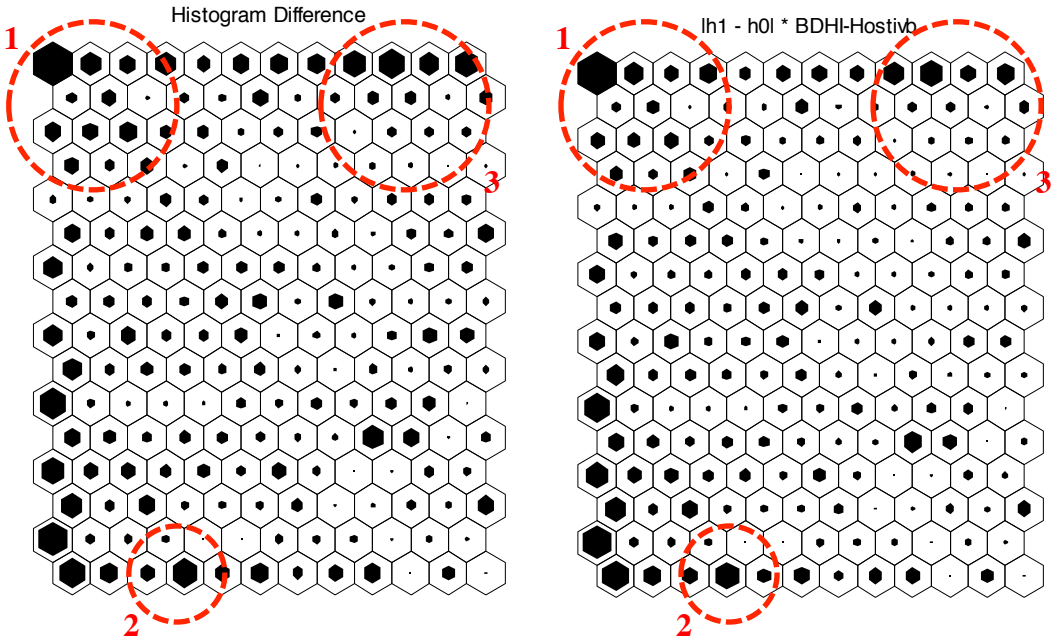


Figura [41] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `BDHI-Hostivb` con inicialización lineal.

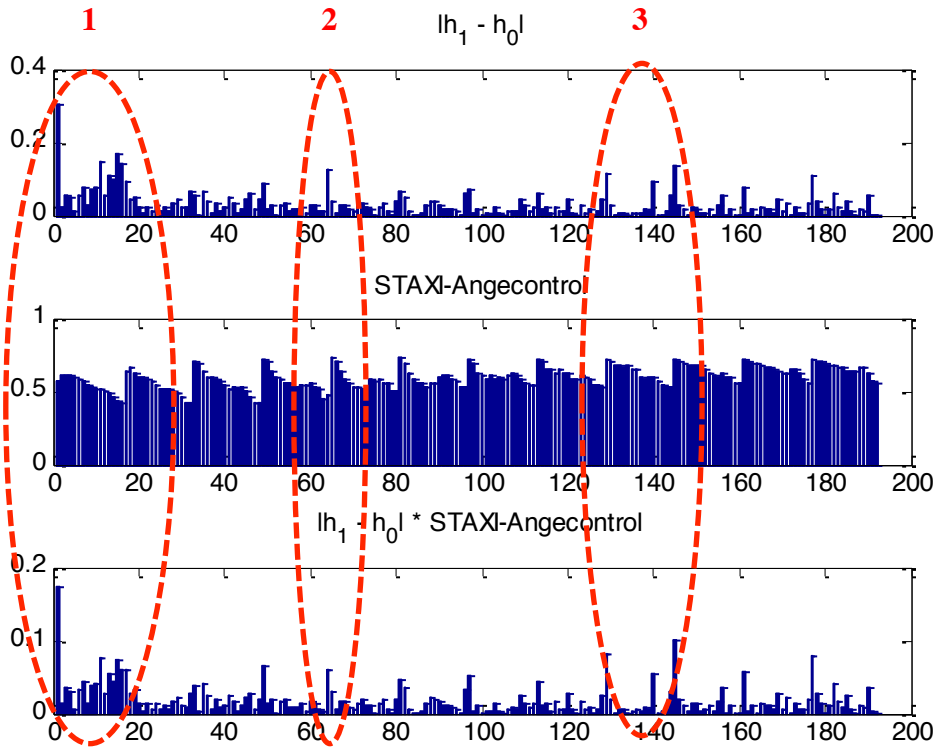


Figura [42] – Diferencia entre histogramas, centroides de la variable `STAXI-Angecontrol` y producto entre ambas gráficas con inicialización lineal.

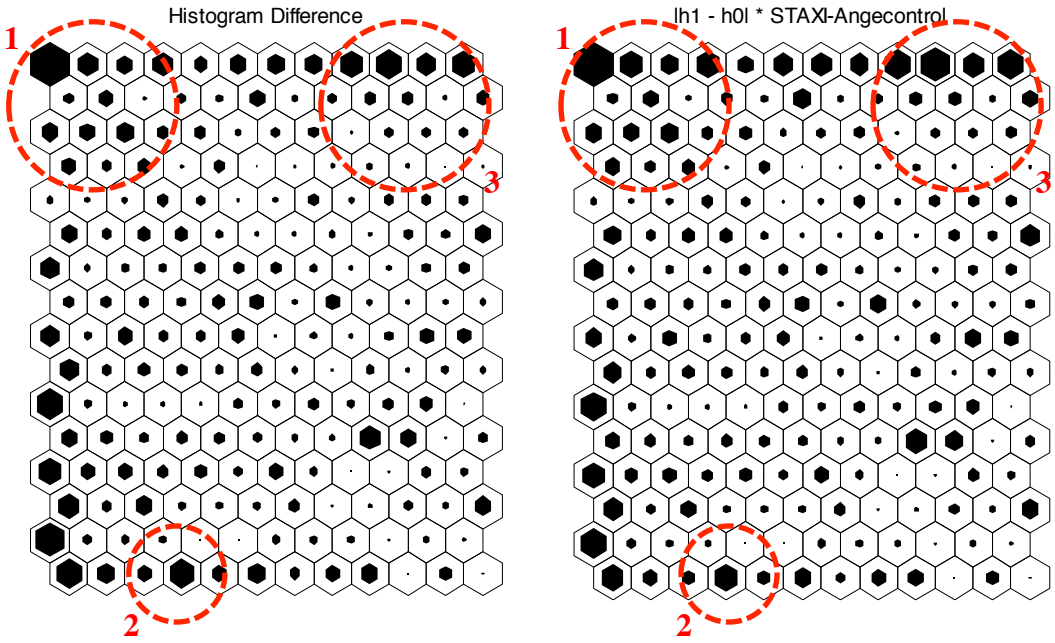


Figura [43] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `STAXI-Angecontrol` con inicialización lineal.

Inicialización con Proyección LDA

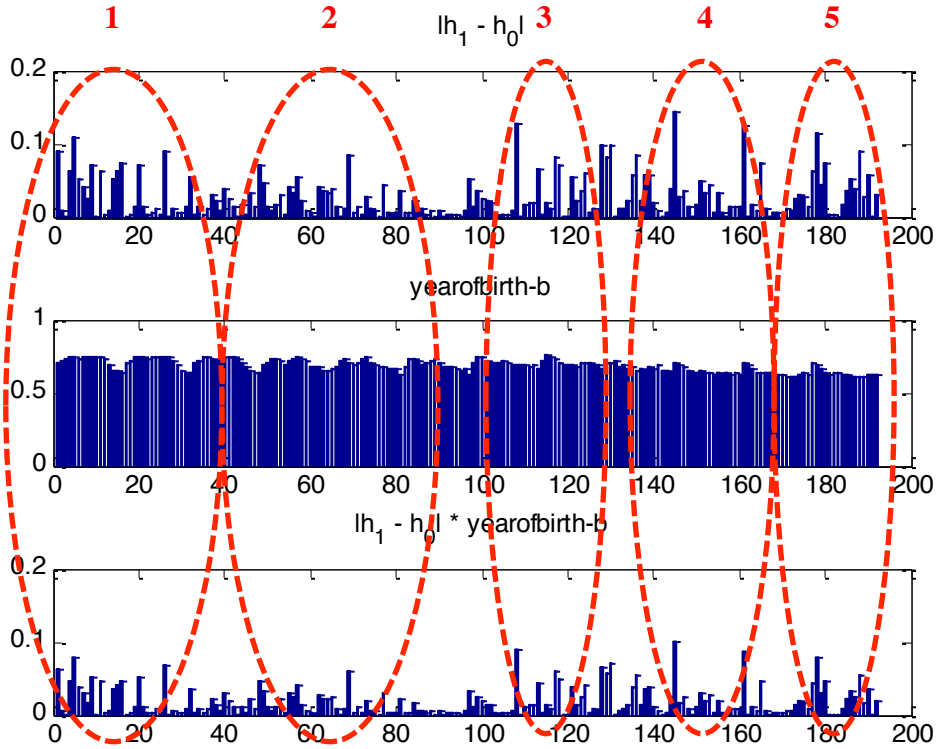


Figura [44] – Diferencia entre histogramas, centroides de la variable `yearofbirth-b` y producto entre ambas gráficas con inicialización LDA.

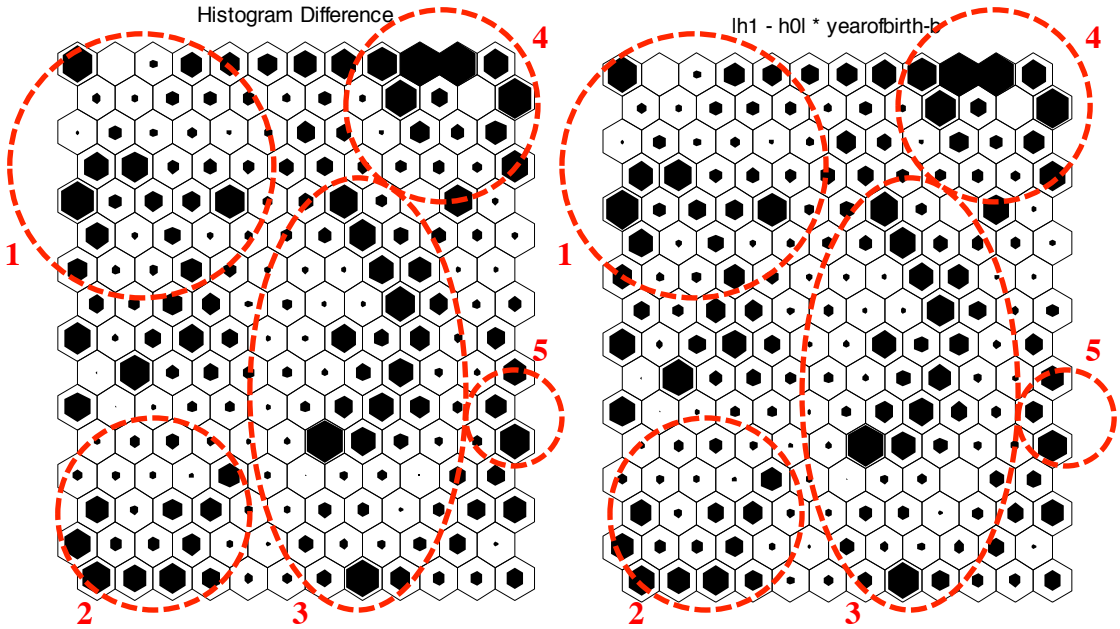


Figura [45] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `yearofbirth-b` con inicialización LDA.

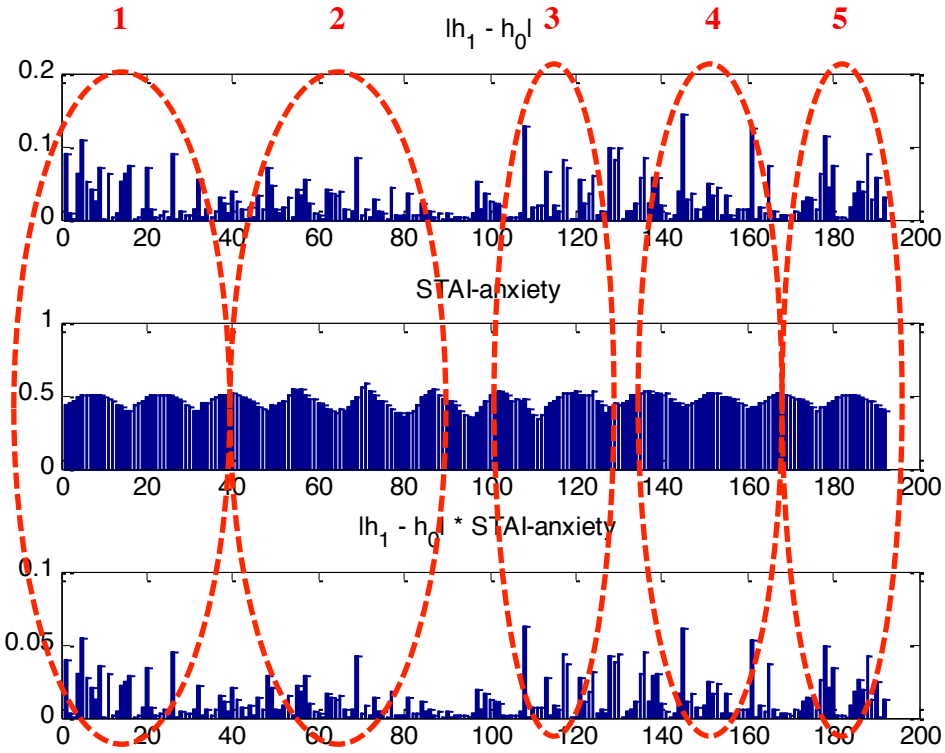


Figura [46] – Diferencia entre histogramas, centroides de la variable `STAI-anxiety` y producto entre ambas gráficas con inicialización LDA.

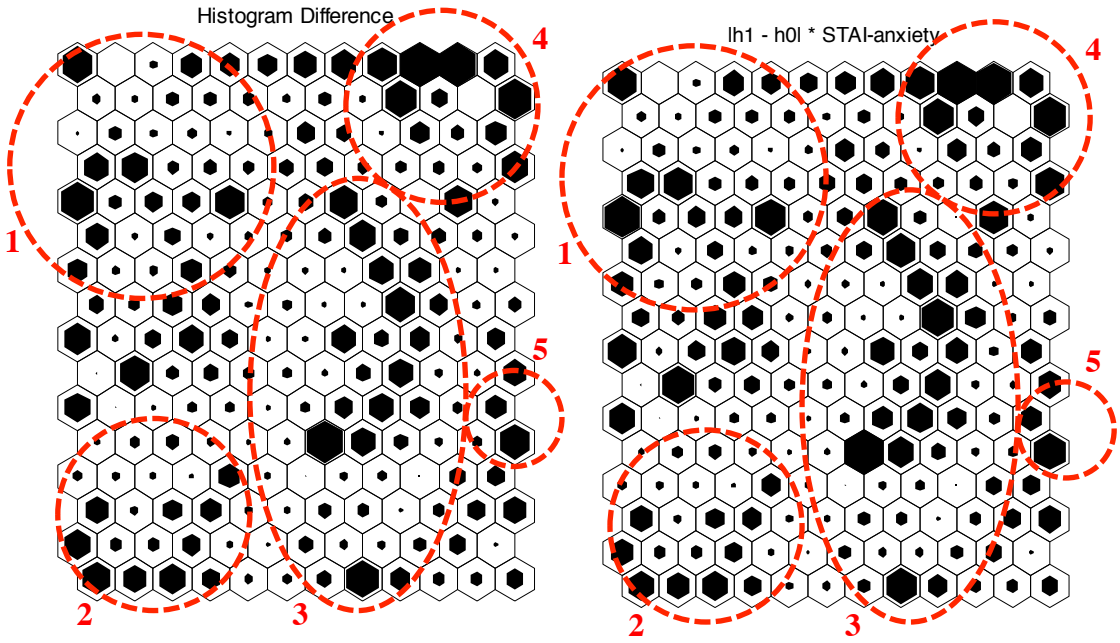


Figura [47] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `STAI-anxiety` con inicialización LDA.

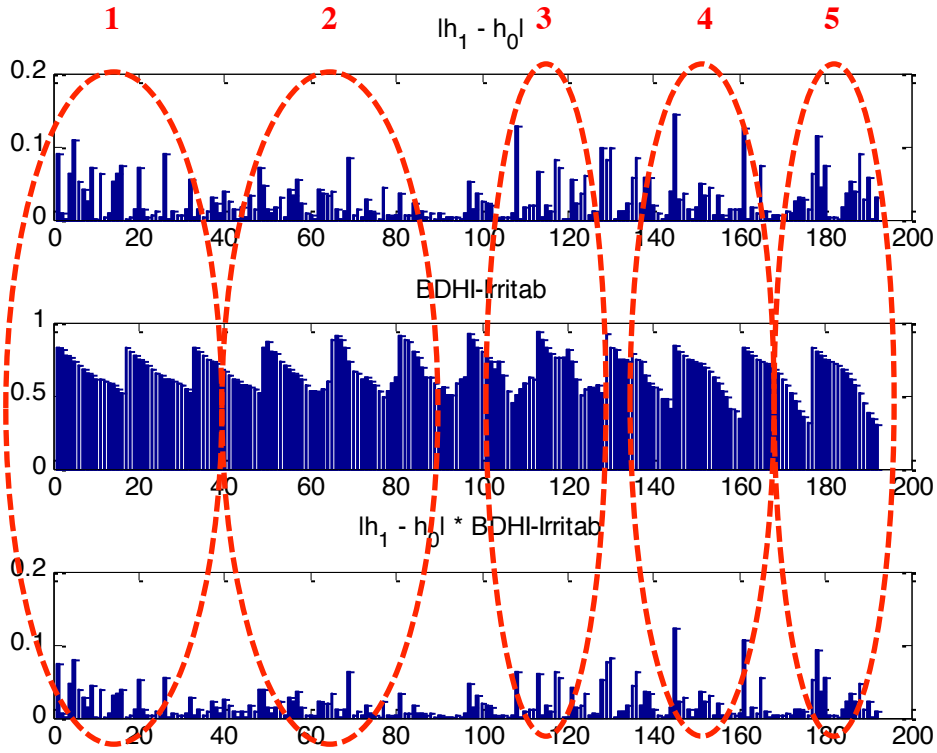


Figura [48] – Diferencia entre histogramas, centroides de la variable `BDHI-Irritab` y producto entre ambas gráficas con inicialización LDA.

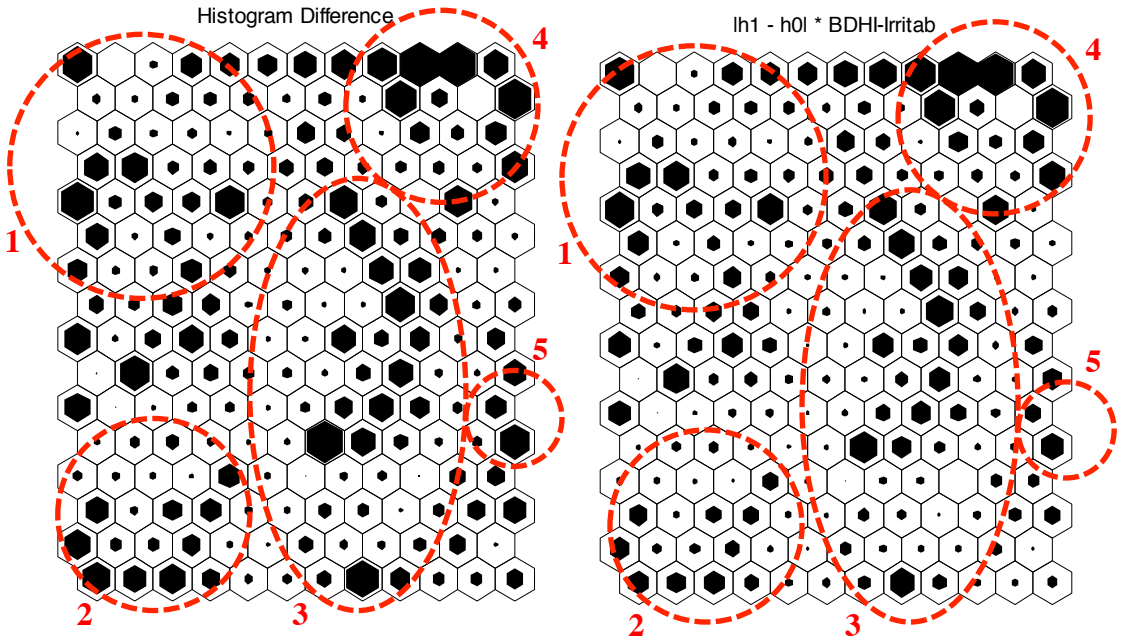


Figura [49] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `BDHI-Irritab` con inicialización LDA.

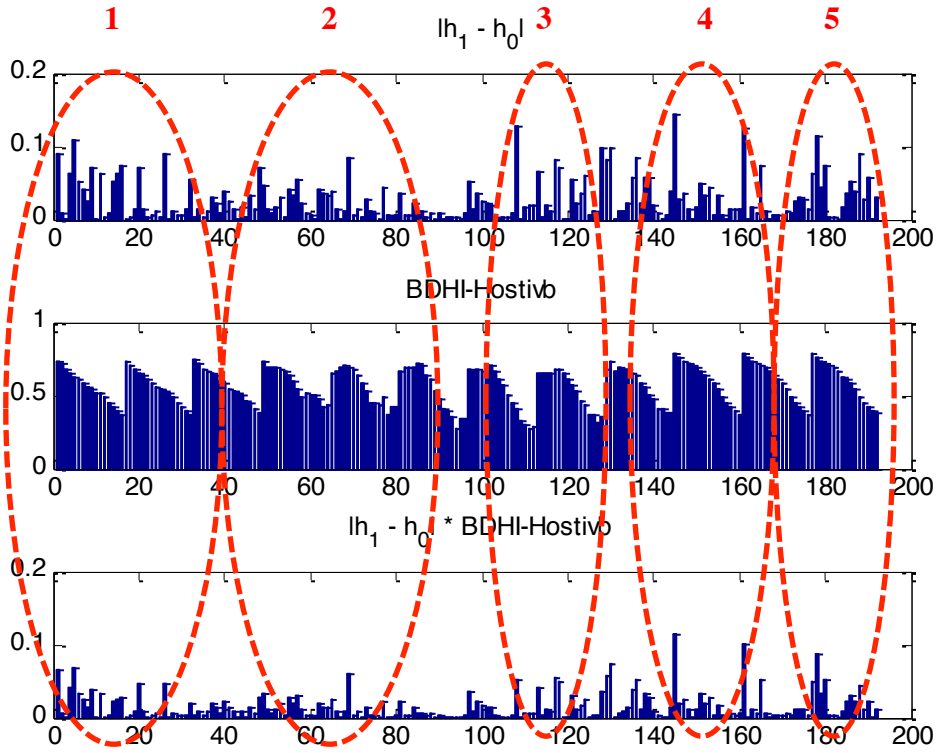


Figura [50] – Diferencia entre histogramas, centroides de la variable `BDHI-Hostivb` y producto entre ambas gráficas con inicialización LDA.

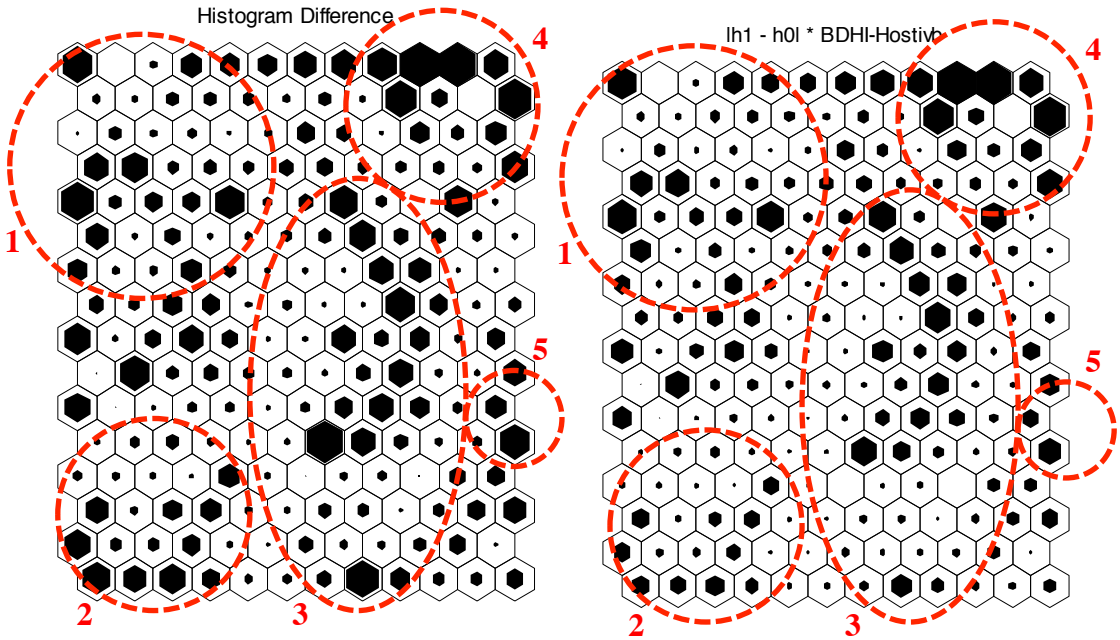


Figura [51] – Histogramas hexagonales de la diferencia entre histogramas y de su producto con el vector de centroides de la variable `BDHI-Hostivb` con inicialización LDA.

3.1.5. Visualización de Gráficas de Variables

En este apartado, se visualizan los mapas autoorganizados correspondientes a las cuatro variables más importantes de acuerdo al criterio de Fisher aplicado a SOM y al discriminante basado en histogramas para los distintos tipos de inicialización. Estas gráficas deberán compararse con la variable suicida para identificar similitudes y diferencias entre zonas frías y calientes.

Discriminante de Fisher aplicado a SOM

En la Tabla [22] se registran los cuatro factores de interés para el discriminante $d_{F_{SOM}}$ estudiado previamente aplicándose un criterio de inicialización aleatoria. La Tabla [23] recoge los resultados para el mismo discriminante utilizando un método de inicialización lineal. Por último, en la Tabla [24] se incluyen las variables más importantes según el discriminante de Fisher aplicado a SOM implementándose una inicialización con proyección LDA.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,2387	0,6623	0,4730	0,4918	0,4391
his_fam_suicide_behavior	0,3518	0,1152	0,3193	0,4495	0,3078
dd_anxiety	0,0848	0,0285	0,1169	0,1245	0,2334
dd_depre	0,0757	0,0325	0,1069	0,1101	0,1995

Tabla [22] – Variables más importantes de acuerdo al discriminante de Fisher aplicado a SOM para inicialización aleatoria.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
gender	0,1313	0,6623	0,4730	0,4918	0,5505
his_fam_suicide_behavior	0,4914	0,1152	0,3193	0,4495	0,4894
dd_oh	0,4349	0,1186	0,3234	0,4412	0,4136
dd_al_drug	0,6380	0,3325	0,4712	0,4791	0,3214

Tabla [23] – Variables más importantes de acuerdo al discriminante de Fisher aplicado a SOM para inicialización lineal.

Variable	$\hat{c}$	μ_0	σ_0	σ_1	$d_{F_{SOM}}$
dd_oh	0,5395	0,1186	0,3234	0,4412	0,5505
EST_CIV_3	0,4313	0,1068	0,3089	0,3541	0,4894
niv_edu	0,6547	0,3746	0,3237	0,3536	0,4136
his_fam_suicide_behavior	0,3623	0,1152	0,3193	0,4495	0,3214

Tabla [24] – Variables más importantes de acuerdo al discriminante de Fisher aplicado a SOM para inicialización LDA.

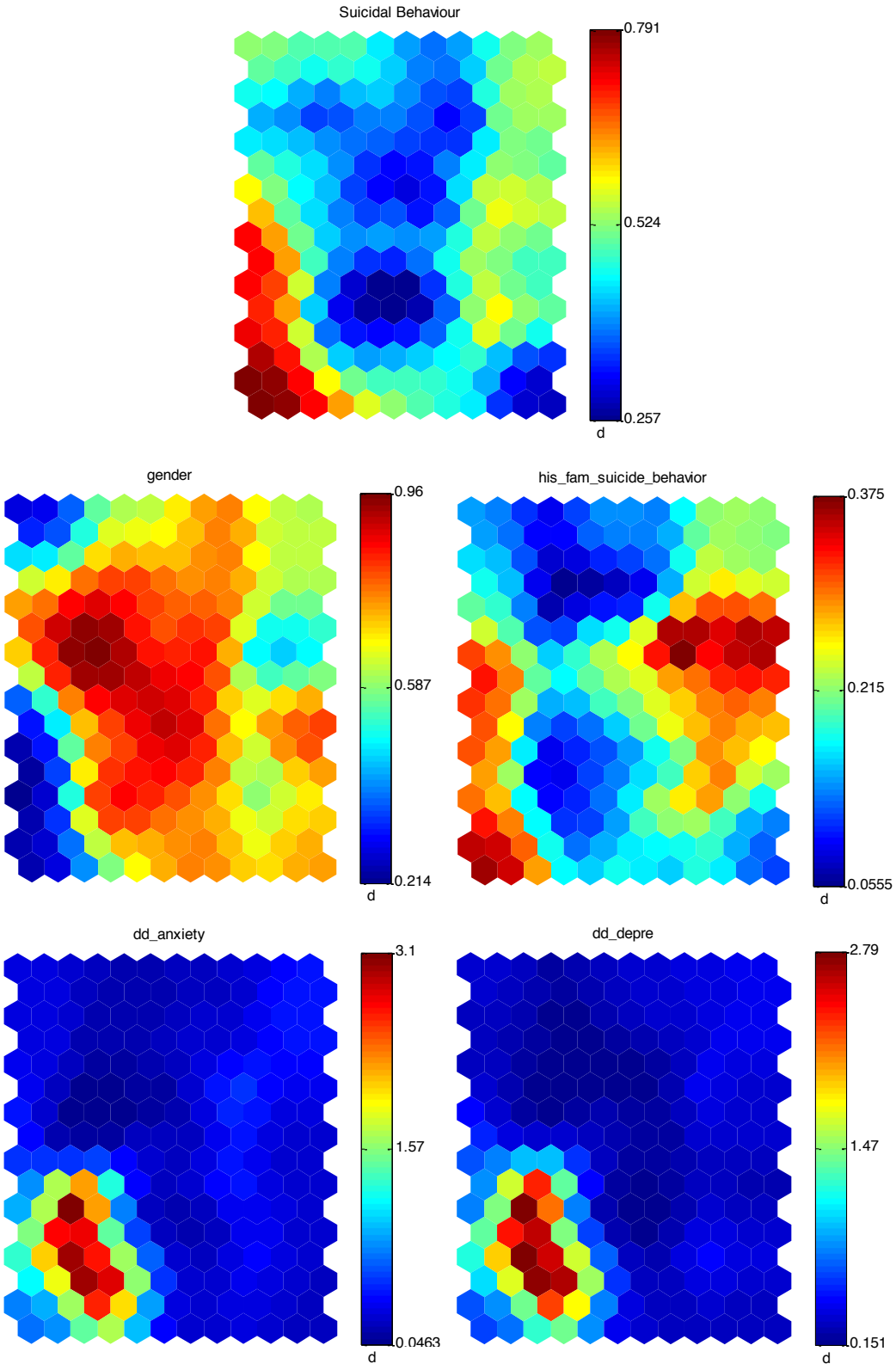


Figura [52] – Variables más importantes según el discriminante $d_{F_{SOM}}$ con inicialización aleatoria.

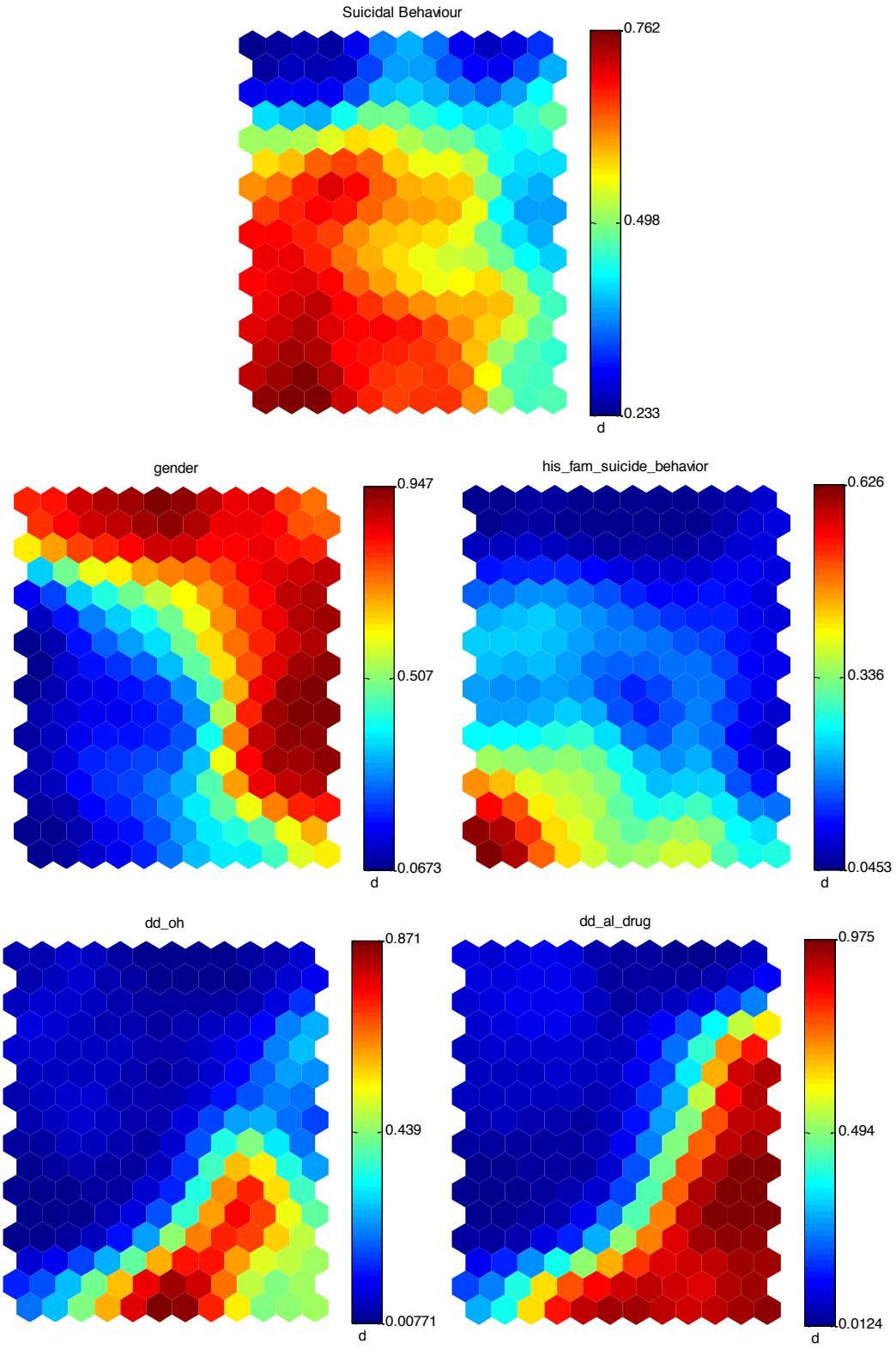


Figura [53] – Variables más importantes según el discriminante $d_{F_{SOM}}$ con inicialización lineal.

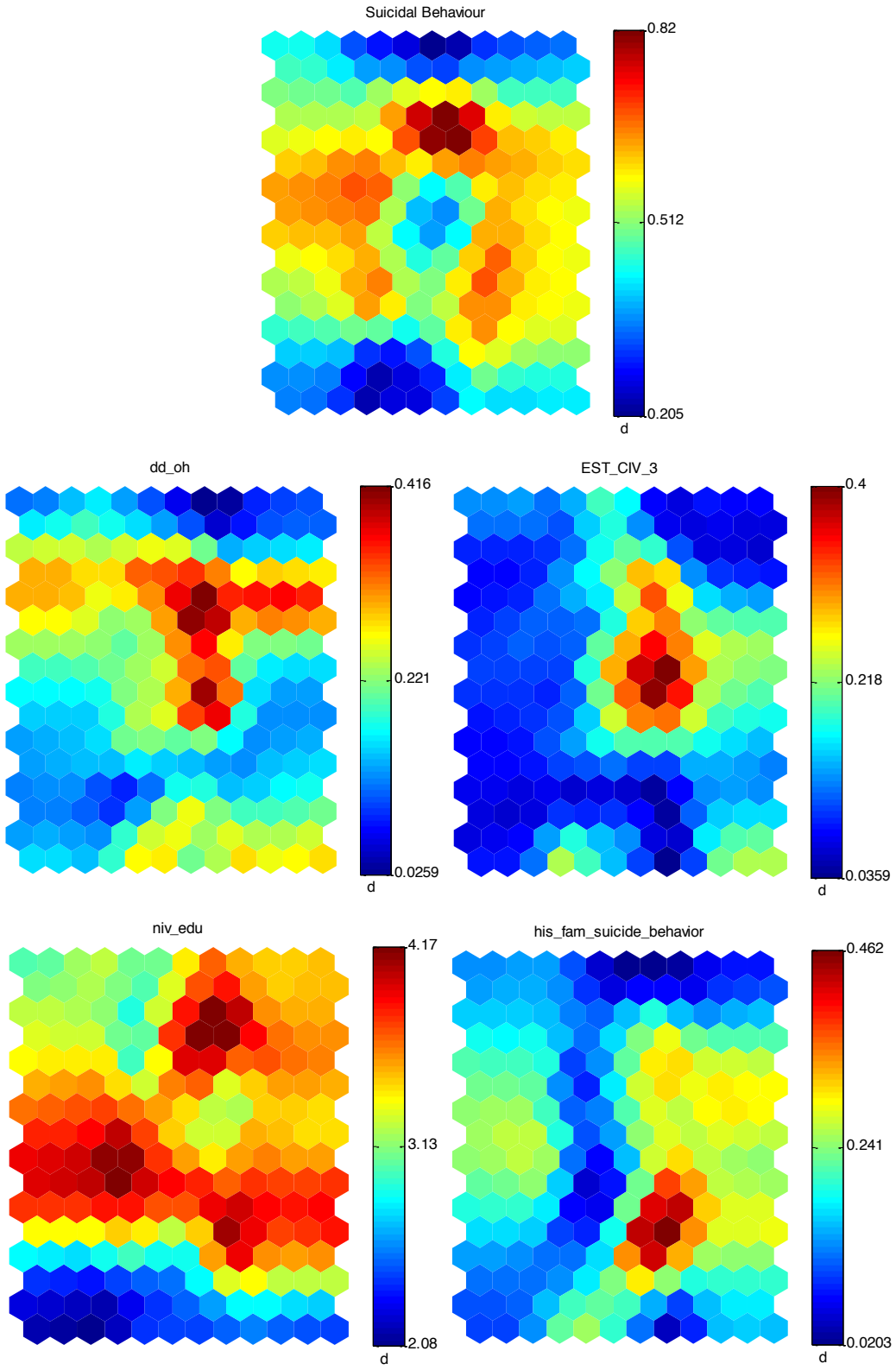


Figura [54] – Variables más importantes según el discriminante $d_{F_{SOM}}$ con inicialización LDA.

En el caso de inicialización aleatoria, las cuatro variables destacadas se corresponden con el género (`gender`), 0 si es hombre y 1 si es mujer, los antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`), la ansiedad (`dd_anxiety`) y la depresión (`dd_depre`).

Con respecto al género, puede apreciarse cómo la zona fría del mapa está correlacionada con la zona caliente de la variable suicida. Esto quiere decir que la probabilidad de suicidio es mayor en mujeres que en hombres. Por otro lado, la correspondencia de algunos de los puntos calientes en antecedentes familiares de conducta suicida con la zona roja del mapa de la variable de comportamiento suicida, indica que, en parte, este factor afecta al desarrollo de posibles intentos de suicidio con desenlace fatal. Si bien, existe otra zona caliente en la variable antecedentes familiares que se solapa con valores intermedios, e incluso fríos, de la variable suicida, por lo que, aunque este factor pueda ser influyente en la conducta suicida, no es, en absoluto, determinante.

Las variables de ansiedad y depresión pueden analizarse conjuntamente debido a sus similitudes gráficas. En ambos casos, la zona de mayor riesgo se sitúa próxima a los puntos calientes de la variable suicida, lo que indica que estos dos factores pueden tener una gran influencia sobre los intentos de suicidio.

En cuanto al método de inicialización lineal, las variables más influyentes son, de nuevo, el género (`gender`) y los antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`), así como el abuso de alcohol (`dd_oh`) y el abuso conjunto de alcohol y drogas (`dd_al_drug`).

En relación al género, se extraen las mismas conclusiones. La correlación de zonas frías de la variable `gender` con zonas calientes de la variable suicida, indica que el riesgo de suicidio es mayor en mujeres que en hombres. Sin embargo, en lo que respecta a la variable de antecedentes familiares de conducta suicida, esta vez parece determinante la influencia de este factor sobre los intentos de suicidio, debido a la coincidencia de puntos calientes en ambas gráficas.

El abuso de alcohol y el abuso conjunto de alcohol y drogas pueden también desencadenar intentos de suicidio. En la zona inferior intermedia de ambos mapas se produce un solapamiento de puntos de color rojo intenso con áreas calientes de la variable de comportamiento suicida, lo que justifica su relevancia para la detección de perfiles de interés para el estudio.

Por último, para el criterio de inicialización con proyección LDA, los factores más importantes son el abuso de alcohol (`dd_oh`), el divorcio (`EST_CIV_3`), el nivel educativo (`niv_edu`) y los antecedentes familiares de conducta suicida (`his_fam_suicide_behavior`).

En cuanto al abuso de alcohol, se aprecia una coincidencia en las zonas de mayor riesgo con respecto a la variable suicida, lo que implica que, a mayor consumo de alcohol, mayor probabilidad de intento de suicidio. La variable `EST_CIV_3` identifica con un 1 a los sujetos separados o divorciados y con un 0 a los que no cumplen con esa condición. El solapamiento de los puntos calientes del mapa con las zonas de color rojo intenso en la variable suicida supone un alto riesgo de suicidio en personas que se encuentren en este estado civil.

Observando la gráfica del nivel educativo, se produce una contradicción con respecto a los conocimientos a priori disponibles. En principio, una persona con estudios de mayor nivel, corre menos riesgo de intento de suicidio que un sujeto sin estudios o con un nivel de educación básico. Sin embargo, la correspondencia entre puntos calientes del mapa de la variable `niv_edu` y el mapa de la conducta suicida indica que perfiles con un nivel alto de estudios tienen una elevada probabilidad de suicidio, mientras que la correlación entre zonas frías en ambos mapas resta riesgo de suicidio en personas con un nivel bajo de estudios.

Finalmente, el solapamiento de zonas calientes de la variable de antecedentes familiares de conducta suicida con áreas de nivel de riesgo intermedio de la variable suicida representa una cierta, pero no determinante, influencia de este factor sobre los intentos de suicidio.

Discriminante basado en Histogramas

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	18,7933
STAI_anxiety	0,1345	14,3216
BDHI_Hostivb	0,1885	13,3905
BIS10_impulnp	0,1581	12,5693

Tabla [25] – Variables más importantes de acuerdo al discriminante basado en histogramas para inicialización aleatoria.

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	26,6727
STAI_anxiety	0,1345	21,6075
BDHI_Hostivb	0,1885	18,1740
STAXI_Angecontrol	0,1986	17,8963

Tabla [26] – Variables más importantes de acuerdo al discriminante basado en histogramas para inicialización lineal.

Variable	σ	d_{Hist}
yearofbirth_b	0,1513	24,0305
STAI_anxiety	0,1345	18,3097
BDHI_Irritab	0,2154	16,6147
BDHI_Hostivb	0,1885	16,4282

Tabla [27] – Variables más importantes de acuerdo al discriminante basado en histogramas para inicialización LDA.

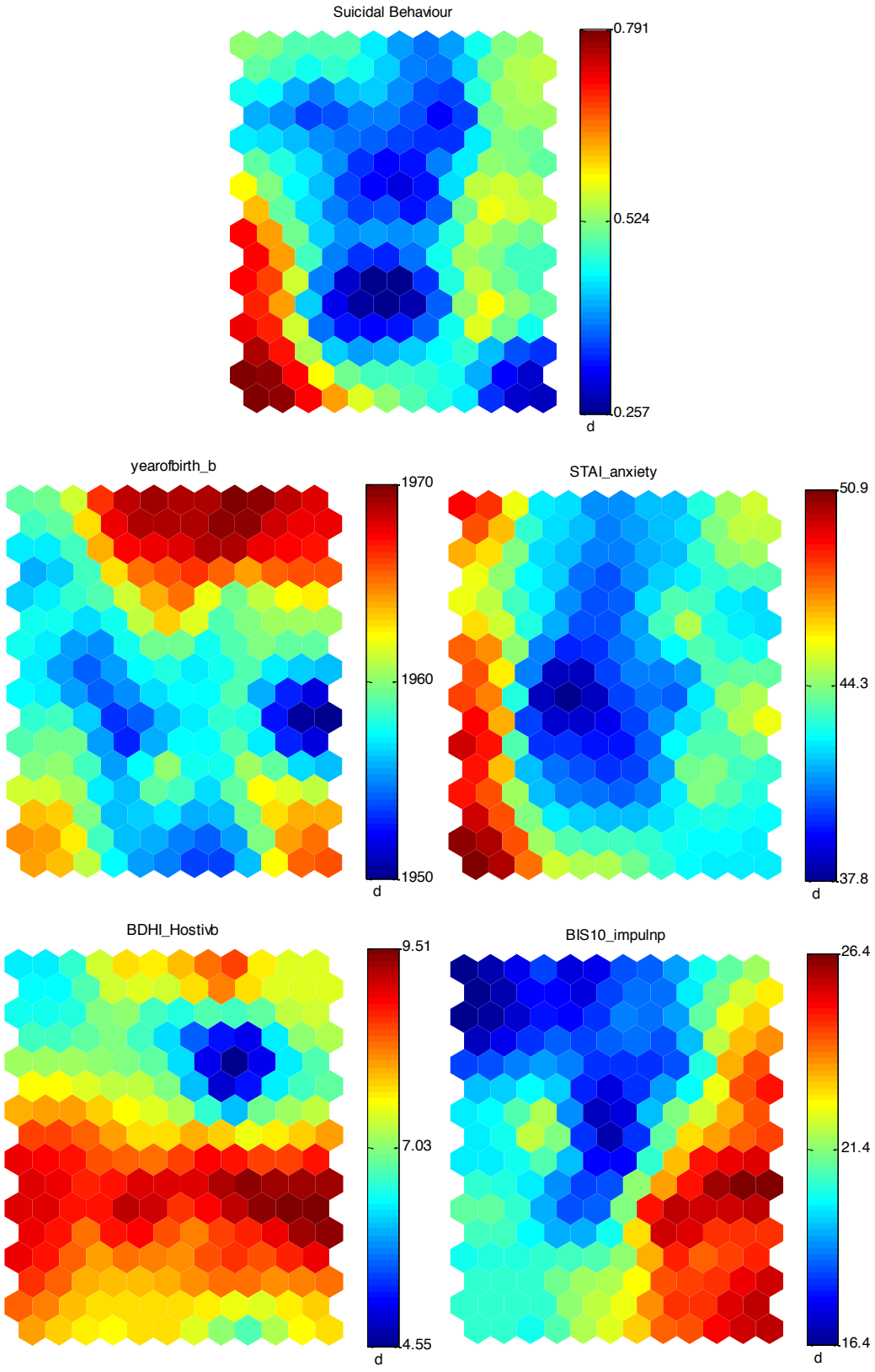


Figura [55] – Variables más importantes según el discriminante d_{Hist} con inicialización aleatoria.

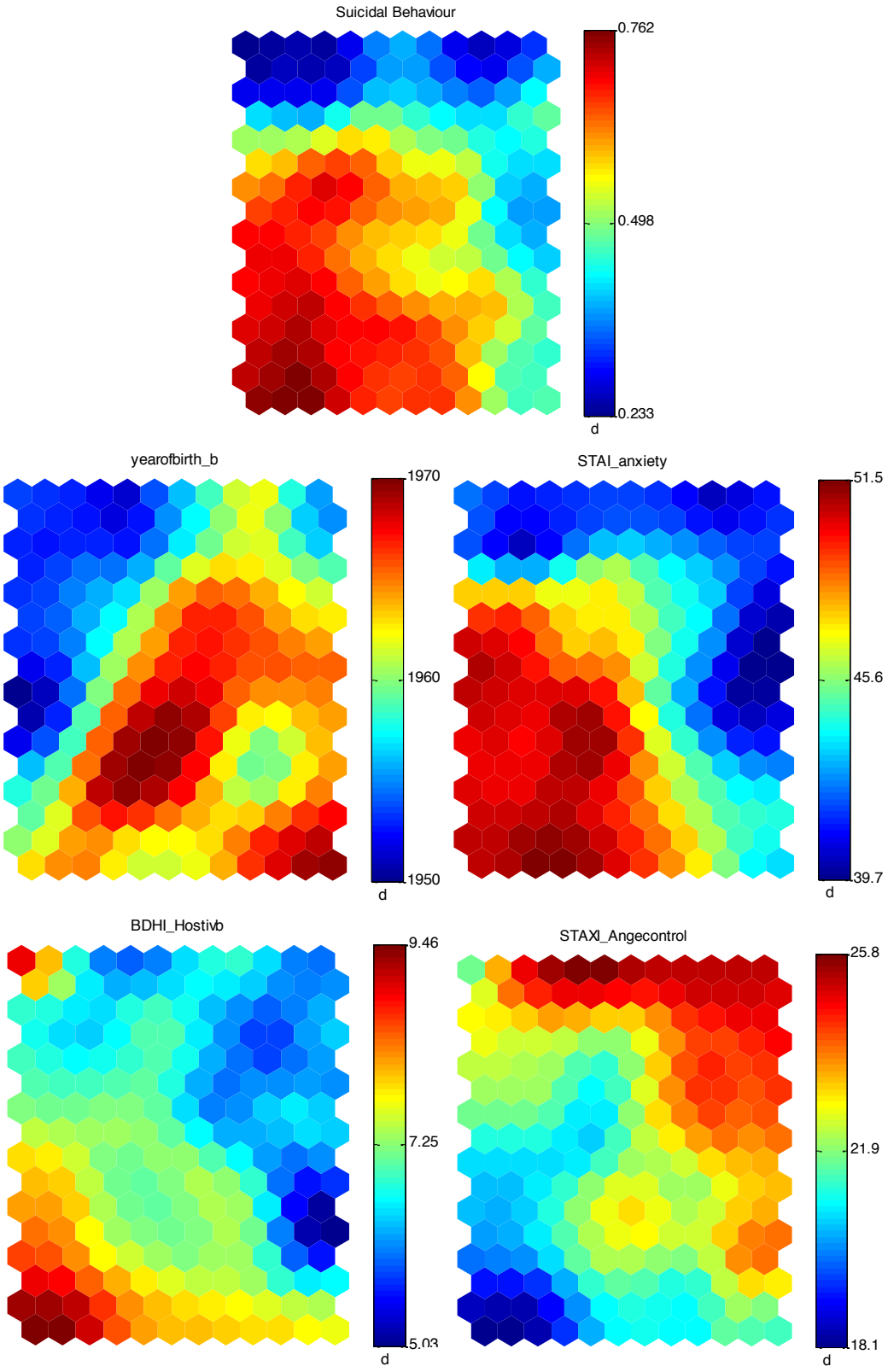


Figura [56] – Variables más importantes según el discriminante d_{Hist} con inicialización lineal.

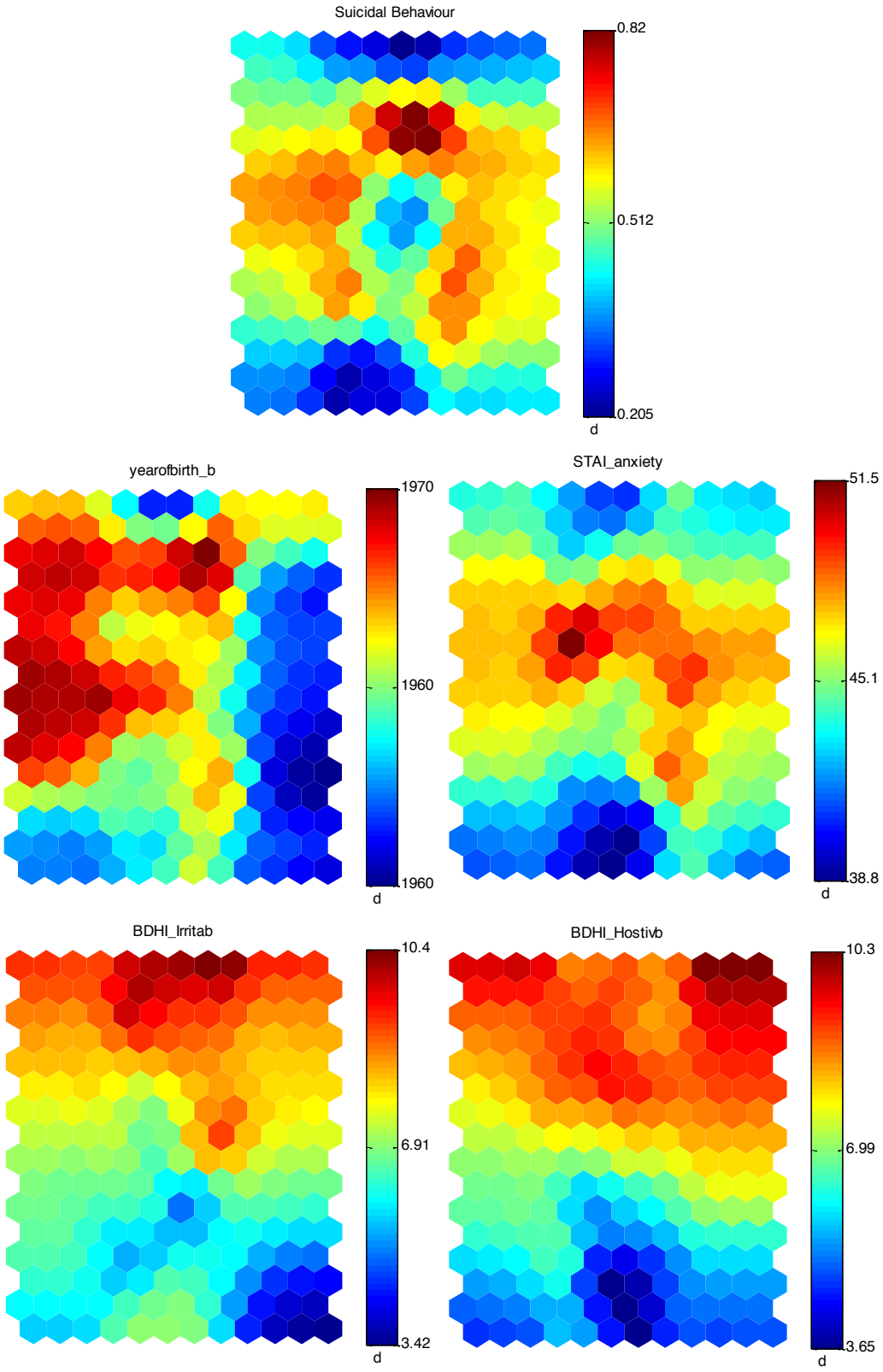


Figura [57] – Variables más importantes según el discriminante d_{Hist} con inicialización LDA.

Analizando los resultados para el discriminante basado en histogramas se tiene que, para inicialización aleatoria, las cuatro variables más importantes se corresponden con el año de nacimiento (`yearofbirth_b`), el estado de ansiedad de acuerdo a los criterios de Molise y Montpellier (`STAI_anxiety`), la hostilidad verbal (BDHI_`Hostivb`) y la impulsividad medida según la planificación (`BIS10_impulnp`).

En el caso del año de nacimiento, existe una correlación de los puntos calientes de la variable suicida con zonas intermedias definidas por sujetos nacidos entorno al año 1963. No obstante, esta variable parece no tener relevancia en la conducta suicida. Por otro lado, bien es cierto que, a mayor ansiedad, mayor probabilidad de suicidio. Esta relación viene dada por la coincidencia entre zonas calientes de la variable suicida y la variable `STAI_anxiety`.

La hostilidad verbal también guarda relación directa con los intentos de suicidio, debido a que las zonas calientes en ambas gráficas se encuentran solapadas. Por último, la impulsividad medida en base a la planificación guarda una correlación inversa con respecto a la conducta suicida. Esto se debe al planteamiento negativo de las preguntas del cuestionario. Es decir, si un sujeto no planifica sus acciones, significa que muestra un perfil impulsivo y, por tanto, más propenso a una conducta suicida. Luego la no planificación, o lo que es lo mismo, las zonas frías del mapa, se corresponden con áreas calientes de la variable suicida.

En lo que respecta a la inicialización lineal, de nuevo la variable año de nacimiento vuelve a ser uno de los factores de riesgo más importantes. Si bien, la conducta suicida está correlacionada con áreas del mapa correspondientes a personas nacidas alrededor del año 1960. Otras variables de interés son el estado de ansiedad de Molise y Montpellier (`STAI_anxiety`), así como la hostilidad verbal (`BDHI_Hostivb`). En ambos casos, es notable el solapamiento de zonas calientes y, por tanto, la fuerte influencia sobre la conducta suicida.

La cuarta y última variable relevante de acuerdo al método de inicialización lineal es el control de enfados (`STAXI_Angecontrol`). De nuevo, existe una correlación inversa con respecto a la variable suicida. El motivo de esta inversión en las zonas frías y calientes es que, cuanto menor control de enfados tenga un sujeto, más impulsivo será y, en consecuencia, mayor será el riesgo de suicidio. Luego, personas con escaso control de enfados, manifestarán una mayor probabilidad de suicidio.

Para finalizar, las variables de interés según la inicialización con proyección LDA son el año de nacimiento (`yearofbirth_b`), el estado de ansiedad de Molise y Montpellier (`STAI_anxiety`), la irritabilidad (`BDHI_Irritab`) y la hostilidad verbal (`BDHI_Hostivb`).

En esta ocasión la variable año de nacimiento se encuentra solapada con zonas calientes del mapa de la variable suicida, por lo que, personas nacidas entorno al año 1970 tendrán mayor riesgo de suicidio. Nuevamente, el estado de ansiedad desencadena posibles intentos de suicidio debido a la correspondencia entre los puntos calientes de ambas gráficas.

En el caso de la irritabilidad y la hostilidad verbal, ambas variables se encuentran fuertemente correlacionadas con la variable suicida. Es decir, las zonas de mayor riesgo se encuentran solapadas con los puntos más calientes de la gráfica `Suicidal Behavior`. Son, por tanto, factores muy influyentes sobre los intentos de suicidio.

3.1.6. Relación entre Discriminantes

A continuación, se muestra la gráfica que representa la relación entre el discriminante de Fisher y el discriminante de Fisher aplicado a SOM. Se espera que la dependencia entre los dos parámetros no sea lineal, si no que aparezcan muestras situadas fuera de la diagonal principal.

La Figura [58] recoge los resultados para un método de inicialización aleatoria. A su vez, en la Figura [59], se incluye la gráfica correspondiente al criterio de inicialización lineal, mientras que la Figura [60] representa la relación de discriminantes aplicando, en este caso, una inicialización con proyección LDA.

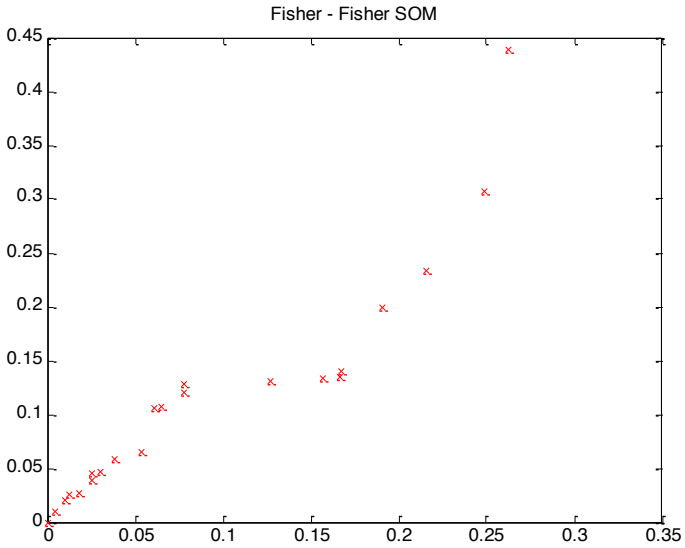


Figura [58] – Relación de discriminantes con inicialización aleatoria.

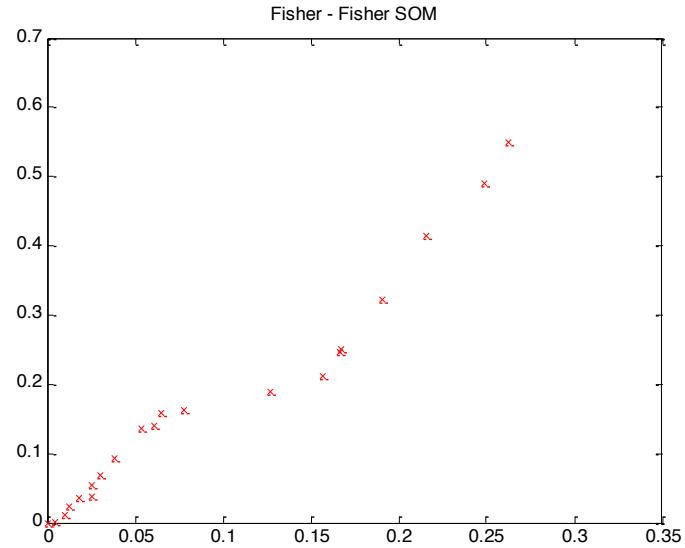


Figura [59] – Relación de discriminantes con inicialización lineal.

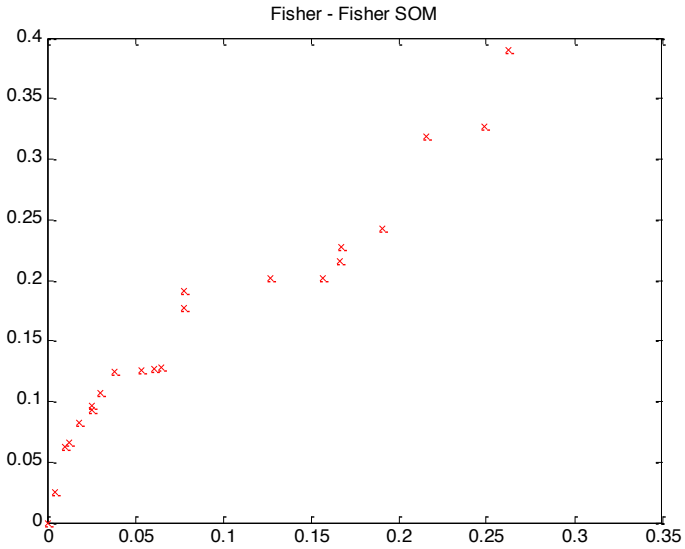


Figura [60] – Relación de discriminantes con inicialización con proyección LDA.

3.2. Conclusiones

En este proyecto se ha aplicado una nueva técnica para la selección de variables basada en los Mapas Autoorganizados. El estudio se ha realizado sobre una cohorte de 8.699 sujetos y 610 variables, incluida la conducta suicida. En la base de datos hay recopilados factores socioeconómicos, psicológicos y clínicos procedentes de un conjunto de individuos entre los que se encuentran perfiles suicidas.

Para la clasificación de variables, se ha recurrido a un criterio discriminativo inspirado en Fisher y a otro método de distinción basado en histogramas. El desarrollo de diferentes tipos de inicialización así como el análisis para distintos tamaños de mapa y porcentajes de restricción, han dotado al proyecto de una mayor visión y diversidad de interpretaciones.

Los resultados obtenidos aportan información adicional a estudios médicos ya realizados sobre pacientes psiquiátricos, facilitando la tarea de detección y prevención de intentos de suicidio. El mapa correspondiente a la variable que representa la conducta suicida presenta una compleja estructura con múltiples picos, que pueden ser interpretados como la existencia de diferentes grupos o subpoblaciones de sujetos con intentos de suicidio.

El discriminante de Fisher aplicado a SOM ha identificado como factor de riesgo los antecedentes familiares de conducta suicida. Otras variables de interés se corresponden con aspectos sociales como el género, el divorcio y el nivel educativo, trastornos mentales como la depresión y la ansiedad y malos hábitos o abusos como el alcohol y las drogas. Por otro lado, el discriminante basado en histogramas revela que el año de nacimiento, el estado de ansiedad y la hostilidad verbal representan factores influyentes en el comportamiento suicida.

Luego, en conjunto, podría decirse que el estudio ha revelado cinco grupos de variables relacionadas con los picos observados en el SOM: factores sociales, antecedentes familiares, trastornos mentales, alcoholismo y drogadicción.

La localización de estas variables no ha sido una tarea fácil. El análisis de Mapas Autoorganizados supone un gran trabajo y debe tenerse especial cuidado con el manejo de los datos de estudio. Precisamente éste ha sido uno de los puntos más complicados en el desarrollo del proyecto. La mayoría de problemas han venido provocados por los valores no identificados en las variables de la base de datos. Mediante la imputación o supresión de estos parámetros se ha logrado solventar estos inconvenientes.

Además, ha sido importante la dicotomización de variables no numéricas para una mejor interpretación de los datos. Por otro lado, la imposición de porcentajes de restricción sobre los datos para dar una mayor fiabilidad al estudio, ha ayudado también en la obtención de resultados más coherentes con respecto a los conocimientos médicos disponibles.

La correcta implementación de los métodos de inicialización y discriminantes también ha jugado un papel clave en el desarrollo del trabajo. Se necesitaba partir de unos mapas que se inicializasen con valores lo más similares posible a los obtenidos tras la fase de entrenamiento. De esta manera, se ahorraría tiempo y se reducirían pasos de ejecución en el programa. Por otro lado, un método discriminativo bien definido ayudaría en la tarea de detección de variables influyentes sobre los intentos de suicidio.

Otros parámetros importantes en el desarrollo del proyecto son los criterios del coseno y de la distancia, que ayudan a decretar cuáles son las dimensiones óptimas de los mapas. Sin embargo, los resultados obtenidos no son concluyentes al respecto, es decir, no se alcanzan soluciones determinantes que no dejen lugar a dudas del tamaño más apropiado para la ejecución del experimento.

Los resultados reportados por este estudio pretenden ayudar en los procesos de identificación y prevención de intentos de suicidio, complementando así el trabajo de psicólogos y psiquiatras. Si bien, las variables que han sido detectadas como factores de riesgo, deberían ser contrastadas por personal médico cualificado, que determinase su influencia o no sobre el comportamiento suicida.

3.3. Líneas Futuras

A pesar del gran trabajo realizado sobre este estudio, existen ciertos factores que podrían mejorarse a medio o largo plazo. Uno de ellos tiene que ver con los criterios de decisión de tamaño óptimo. De acuerdo a los resultados obtenidos, no podría determinarse qué dimensiones son las más acertadas según qué método de inicialización se aplique.

El criterio basado en la distancia depende implícitamente del número de celdas que integran el mapa, de modo que, a mayores dimensiones, menor distancia entre histogramas. Se necesita aplicar un método independiente del tamaño, que permanezca invariable ante cambios dimensionales.

Con el criterio del coseno sucede algo parecido. El número de celdas determina, en parte, el valor del coseno del ángulo que forman los vectores definidos por los histogramas de muestras positivas y negativas. No se ha encontrado ninguna solución a este problema, por lo que, una posible mejora atañería a estos factores.

Otro aspecto a tener en cuenta es el discriminante basado en histogramas. Aunque los resultados obtenidos tienen cierta coherencia, la presencia de la variable año de nacimiento no es del todo convincente. Para mejorar la detección de factores de interés, podría optimizarse el método de selección de perfiles de centroides. Es decir, debería diseñarse un nuevo proceso que identificase de manera más eficaz vectores de centroides con una distribución parecida a la que sigue la distancia entre histogramas.

Por último, una mejora más a largo plazo tendría que ver con la interfaz del proyecto. La entrada de datos por teclado que se ha implementado resulta bastante rudimentaria si lo que se quiere es hacer de este programa una herramienta médica que sirva de soporte a personal psicológico y psiquiátrico.

La implementación de una nueva interfaz de usuario con una apariencia más visual y atractiva, facilitaría su uso por parte de médicos, investigadores o profesores, haciendo de esta aplicación una herramienta útil en la tarea de detección de intentos de suicidio. El análisis en tiempo real de perfiles de sujetos con alto riesgo de suicidio, ayudaría a determinar las intenciones y probable desarrollo de conducta de estos pacientes.

Si bien, las personas y sus acciones no pueden preverse en su totalidad, por lo que la fiabilidad del programa no sería de un 100%, pero sería un soporte o punto de vista adicional a los métodos utilizados en la actualidad.

La nueva interfaz integraría un menú de selección de parámetros tales como el tipo de inicialización, las dimensiones del mapa o el criterio discriminante aplicado. Además, tendría una base de datos accesible para que el usuario pudiese leer y escribir nuevas variables.

4. Planificación y Presupuesto

En los siguientes apartados se detallará la planificación seguida en la consecución del proyecto así como el presupuesto necesario en personal, hardware y software.

4.1. Planificación

Con el objetivo de completar cada una de las tareas que componen el proyecto, ha sido necesario distribuir el tiempo disponible optimizando así la consecución de los diferentes hitos. Dada la magnitud y múltiples ejecuciones y pruebas del estudio, se han definido cuatro procesos generales que recogen las subtareas programadas para el desarrollo del trabajo.

1. Análisis:
   Esta tarea comprende tres hitos relacionados con una primera toma de contacto del estudio. En primer lugar, se llevará a cabo un estudio de la propuesta, analizando el interés, la complejidad y el posible resultado final. Una vez aprobado el proyecto, comenzará un proceso de documentación para adquirir nociones sobre Mapas Autoorganizados y conocer cuál es el marco social y económico de los intentos de suicidio. Por último, será necesario aprender a manejar la herramienta de desarrollo SOM Toolbox para la implementación de código.

2. Implementación
   Es el proceso de mayor trabajo aunque no de mayor duración. A lo largo de este hito se implementará la totalidad del programa, es decir, la lectura y tratamiento de datos, la inicialización y dimensiones de mapas, los métodos discriminantes, los criterios de selección de tamaño, los porcentajes de restricción, la visualización de gráficas y la interfaz de usuario.

   En esta tarea es donde se han encontrado la gran mayoría de dificultades y ha supuesto el esfuerzo más importante para el desarrollo y consecución del proyecto. Ha sido un proceso de prueba y error que ha determinado el éxito y buen término del estudio.

3. Pruebas
   Este hito comprende el proceso de pruebas y experimentos para analizar los resultados obtenidos, llevar a cabo la toma de decisiones y verificar el correcto funcionamiento del programa. De este proceso se sacan conclusiones positivas y negativas con respecto a cómo se ha ejecutado la implementación del código.

4. Memoria
   Es la tarea de mayor extensión temporal, ya que se inicia desde el proceso de documentación y finaliza una vez se ha implementado el programa y analizado los resultados. Es, sin duda, la visión más completa de lo que ha sido el trabajo y tiempo dedicado al proyecto.

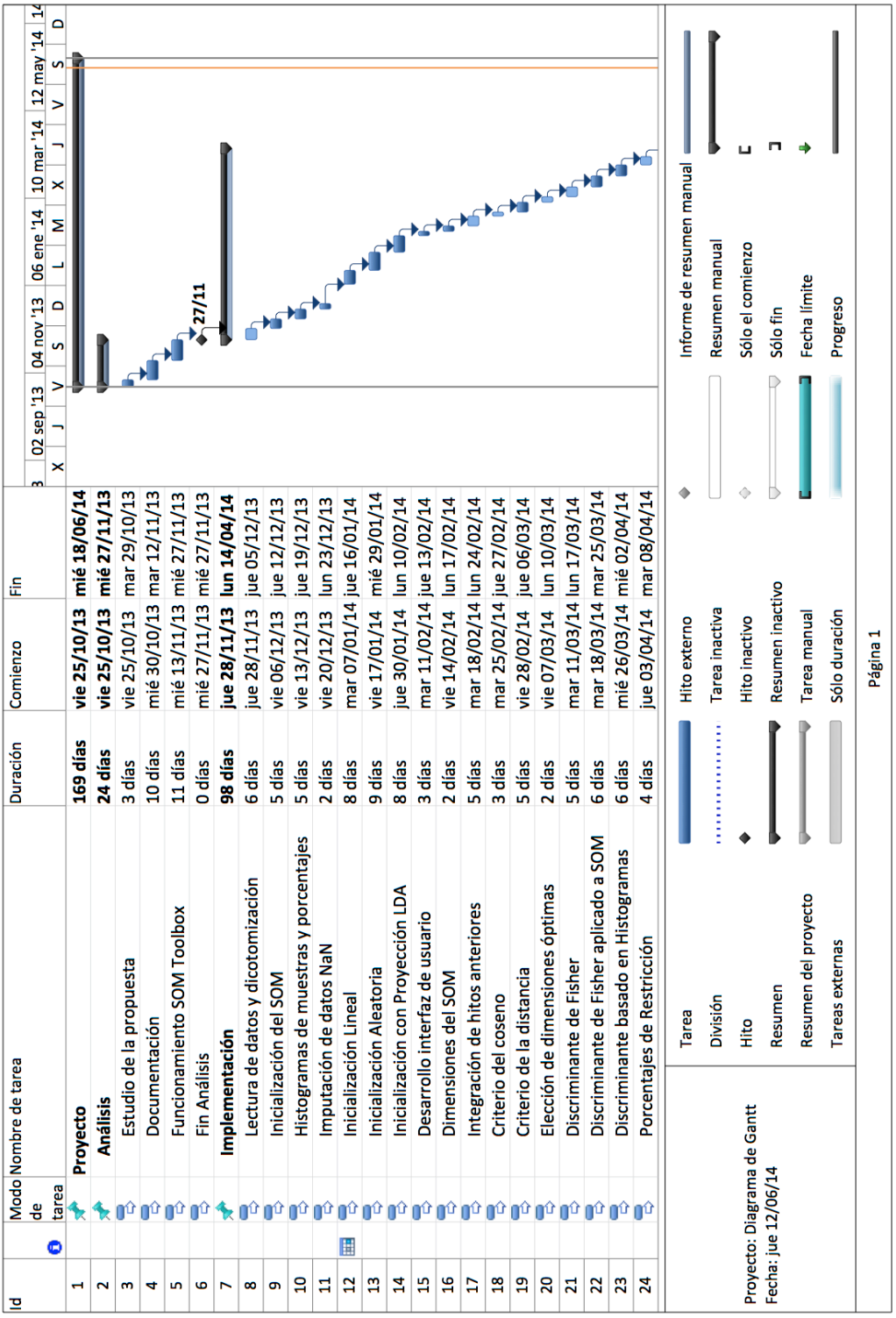


Figura [61] – Diagrama de Gantt 1/2.

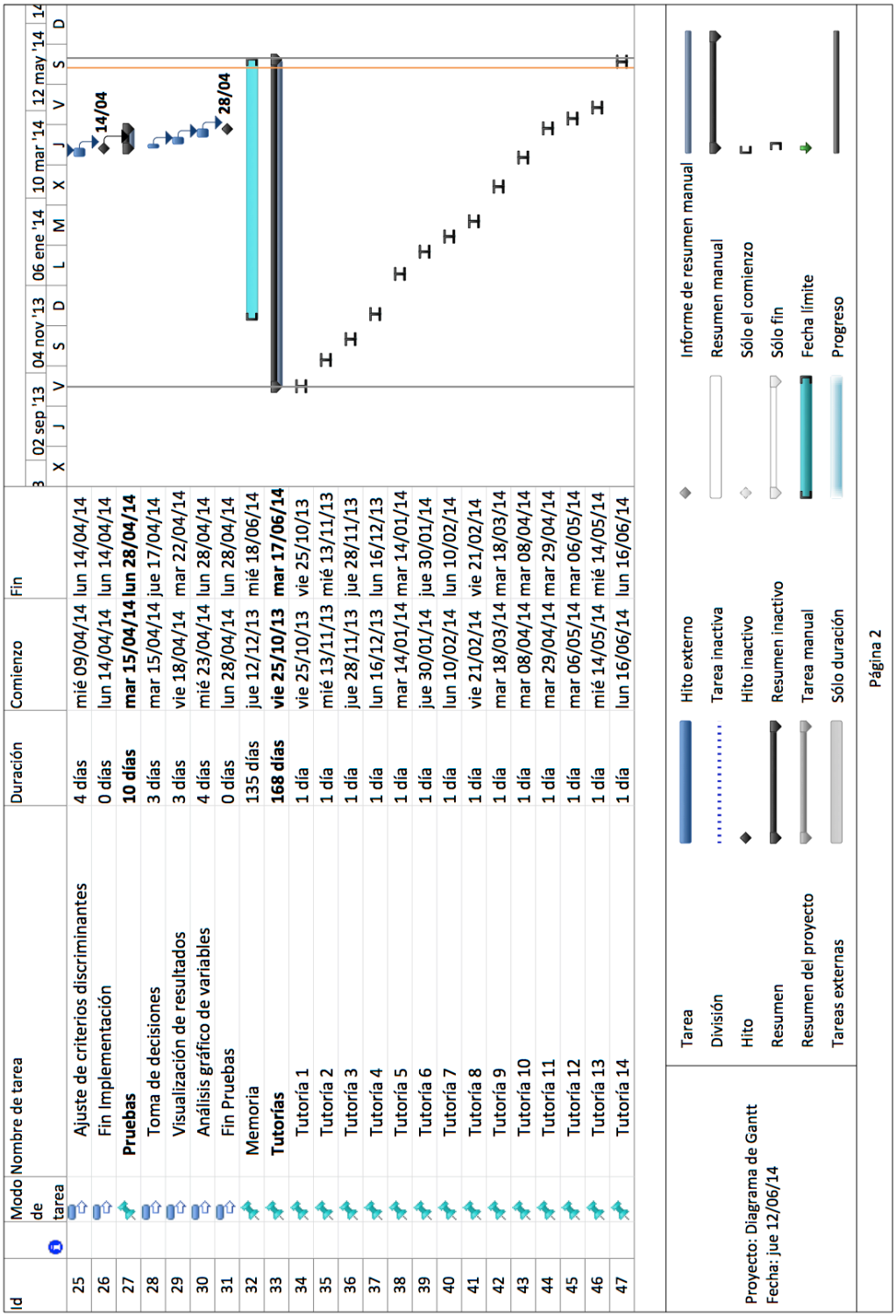


Figura [62] – Diagrama de Gantt 2/2.

En las Figuras [61] y [62] se incluye el Diagrama de Gantt, que representa las tareas que integran el estudio, así como su duración y relación. Las tutorías se han definido como hitos independientes de las fases de desarrollo del proyecto.

4.2. Presupuesto

En este apartado se van a presupuestar todos los recursos materiales y personales que han intervenido en el desarrollo del proyecto. Para determinar el número de horas invertidas en el trabajo, se ha planificado un calendario especificado en la Tabla [28] que incluye el total de horas dedicadas a cada una de las tareas que integran el estudio completo.

Tarea	Días	Horas / Día	Total Horas
Análisis	24	3	72
Implementación	98	4	392
Pruebas	10	3	30
Memoria	135	2	270
TOTAL			**764**

Tabla [28] – Total de horas trabajadas en el desarrollo del proyecto.

Por tanto, el tiempo en días empleado para el desarrollo del proyecto asciende a 764 días. A continuación, se especifica el presupuesto correspondiente al personal, hardware y software necesarios para la consecución del trabajo.

En cuanto al presupuesto de personal, la cuantía percibida por hora se ha extraído de la plantilla presupuestaria proporcionada por la Universidad Carlos III de Madrid, de modo que el sueldo estipulado para cada empleado queda indicado en la Tabla [29].

Personal	Total Horas	Coste / Hora [€]	Total [€]
Analista	72	33	2.376
Programador	392	25	9.800
Responsable de Pruebas	30	15	450
Responsable de Memoria	270	15	4.050
TOTAL			**16.676**

Tabla [29] – Coste total de personal.

Se adjunta también el presupuesto relativo a los materiales hardware y software que han sido necesarios en la implementación del proyecto. Cada uno de los costes queda desglosado en las Tablas [30] y [31].

Con respecto a los recursos hardware será necesario calcular el valor de la amortización para cada dispositivo aplicando la expresión de la Ecuación [19].

$$Total = \frac{A}{B} \; x \; C \; x \; D \qquad [19]$$

donde:

- ✓ A: meses de uso del equipo.
- ✓ B: periodo de depreciación.
- ✓ C: coste del equipo.
- ✓ D: porcentaje de uso dedicado al equipo (normalmente 100%).

Equipo	Cantidad	Coste / Unidad [€]	Uso [%]	Dedicación [meses]	Periodo Depreciación	Total [€]
MacBook Pro 13''	1	1229	100	8	60	163,87
Magic Mouse Apple	1	69	100	8	60	9,2
PC Portátil HP G62	1	699	100	8	60	93,2
TOTAL						**266,27**

Tabla [30] – Coste total hardware.

Aplicación	Licencias	Coste / Unidad [€]	Total [€]
Matlab 2011 Windows	1	174	174
Matlab 2011 Mac	1	174	174
Paquete Office Professional PC 2013	1	399,99	399,99
Paquete Office Home&Student Mac 2011	1	199,99	199,99
Paquete Project Standard 2013	1	589,99	589,99
TOTAL			**1.537,97**

Tabla [31] – Coste total software.

Por tanto, el resumen total de gastos sobre el que se ha aplicado un impuesto de costes indirectos del 20% queda especificado en la Tabla [32].

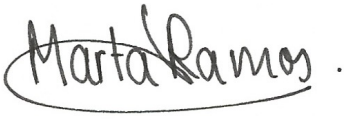
Concepto	Total [€]
Personal	16.676
Hardware	266,27
Software	1.537,97
TOTAL (+20%)	**18.480,24 (22.176,29)**

Tabla [32] – Resumen de costes.

Luego:

"El presupuesto total del proyecto asciende a la cantidad de **VEINTIDOS MIL CIENTO SETENTA Y SEIS CON VEINTINUEVE** euros".

Leganés, a 22 de Junio de 2014



Marta Ramos Martín

5. Bibliografía

[1] Associated Press, 2013. Aumentan los suicidios en el mundo por la crisis económica. Tercera Información.

[2] Baca-García, E., Vaquero-Lorenzo, C., Pérez-Rodríguez, M., Gatacós, M., Bayés, M., Santiago-Mozos, R., Leiva-Murillo, J. M., Prado-Cumplido, M., Artés-Rodríguez, A., Ceverino, A., Díaz-Sastre, C., Fernández-Navarro, P., Costas, J., Fernández-Piqueras, J., Díaz-Hernández, M., León, J., Baca-Baldomero, E., Saiz-Ruiz, J., J. John Mann, Ramin V. Parsey, Carracedo, A., Estivill, X., Oquendo, M.A., 2009. Nucleotide Variation in Central Nervous System Genes Among Male Suicide Attempters.

[3] Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

[4] Investigación y Ciencia, 2012. La plasticidad sináptica, base del aprendizaje y la memoria.

[5] Jieping Ye, Ravi Janardan, Qi Li, 2004. Two-dimensional linear discriminant analysis.

[6] Kohonen, T., 1982. Self-organized formation of topologically correct feature maps.

[7] Kohonen, T., 2001. Self-Organizing Maps, 3rd Edition. Springer, Berlin.

[8] Kohonen, T., Honkela, T., 2007. Kohonen network. Scholarpedia, 2(1):1568.

[9] Leiva-Murillo, J., López-Castromán, J., Baca-García, E., European Research Consortium for Suicide (EURECA), 2012. Characterization of Suicidal Behaviour with Self-Organizing Maps.

[10] López-Castromán, J., Pérez-Rodríguez, M., Jaussent, I., Alegría, A., Artés-Rodríguez, A., Freed, P., Guillaume, S., Jollant, F., Leiva-Murillo, J., Malafosse, A., Oquendo, M., de Prado-Cumplido, M., Saiz-Ruiz, J., Baca-García, E., Courtet, P., European Research Consortium for Suicide (EURECA), 2009. Distinguishing the relevant features of frequent suicide attempters. Journal of Psychiatric Research 45 (5), 619–625.

[11] Massot, M., 2014. Mapa de suicidios desde el inicio de la crisis. El Periódico.

[12] Mengual, E., 2011. Suicidios, la epidemia del siglo XXI. El Mundo.

[13] Ruiz, F. J. R., Valera, I., Blanco, C., Pérez Cruz, F., 2012. Bayesian Nonparametric Modeling of Suicide Attempts.

[14] San-Martín, O., 2014. Los suicidios suben un 11% y se sitúan en su tasa más alta desde 2005. El Mundo.

[15] Serrano, M. M., 2005. La protección de los datos sanitarios. La historia clínica.

[16] Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., Team, S., Oy, L., 2000. Som toolbox for matlab. Techn. Ber., Helsinki University of Technology.

[17] World Health Organization (WHO). Figures and facts about suicide; 1999. Geneva.

[18] World Health Organization (WHO). Suicide Prevention; 2012. http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/